INVERSE PROBLEMS IN HIGH DIMENSIONAL STOCHASTIC SYSTEMS UNDER UNCERTAINTY

by

Patrick Lloyd Harrington Jr.

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2010

Doctoral Committee:

Professor Alfred Hero, Chair
Professor Daniel Burns
Professor Margit Burmeister
Associate Professor Kerby Shedden
Associate Professor Ji Zhu

To My Parents,

Patrick and Cynthia.

If I can attain half of the success
you have achieved in marriage and in life,
I will have lived a full and purposeful life.

A son could not ask for better parents.

ACKNOWLEDGEMENTS

I am extremely grateful to have been advised by a brilliantly creative human being. Professor Alfred Hero has allowed me to mature as an independent researcher capable of abstractly analyzing complex problems. Between day to day interactions, coursework, and discussions about research, I am forever grateful for the interactions I have had with my committee members Professors Burns, Burmeister, Shedden, and Zhu. I will always appreciate the hands on interaction and development of ideas with my post-doctoral researchers Mark Kliger and Ami Wiesel. I am so appreciative of the time and effort you both spent with me, especially early on in my graduate student career. I cannot thank Arvind Rao enough for his wisdom and insight into developing good research topics and sharing a few good laughs on our road trip to Madison, WI. I know he will be a very successful faculty member some day. For both academic collaboration and extracurricular mischief, I will never forget the moments spent with fellow graduate students and now life long friends Yongsheng Huang and Arnau Tibau Puig. Most importantly, I thank my beautiful wife and best friend Erica for her love and patience over these past four years while pursuing my Ph.D.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

1.1 Motivation

The four methods that comprise this thesis are an attempt to identify and solve fundamental components necessary for *Predictive Health and Disease* (PHD). The idea of PHD was originally motivated in a Defense Advanced Research Projects Agency (DARPA) grant where the goal is to develop a socio-molecular inference engine. Such an inference engine would be designed for a cohort of individuals in which an infectious disease can propagate across this population with interactions defined by their social network. Partial molecular and categorial information of individuals disease states and social network topology would be used to update ones posterior probability of all individuals' hidden disease states in the network and select the next measurements which would elicit the largest gain in information regarding these states. The fusion of data from individual high-dimensional biomedical measurements with knowledge of social interactions and the epidemiological of the infectious agent are necessary to develop such an engine. Figure 1.1 summarizes different components within such a PHD socio-molecular inference engine. The interdisciplinary scope of developing a PHD inference engine bridges domains from public health, mathematical epidemiology, bioinformatics, clinical decision making,

and machine learning.

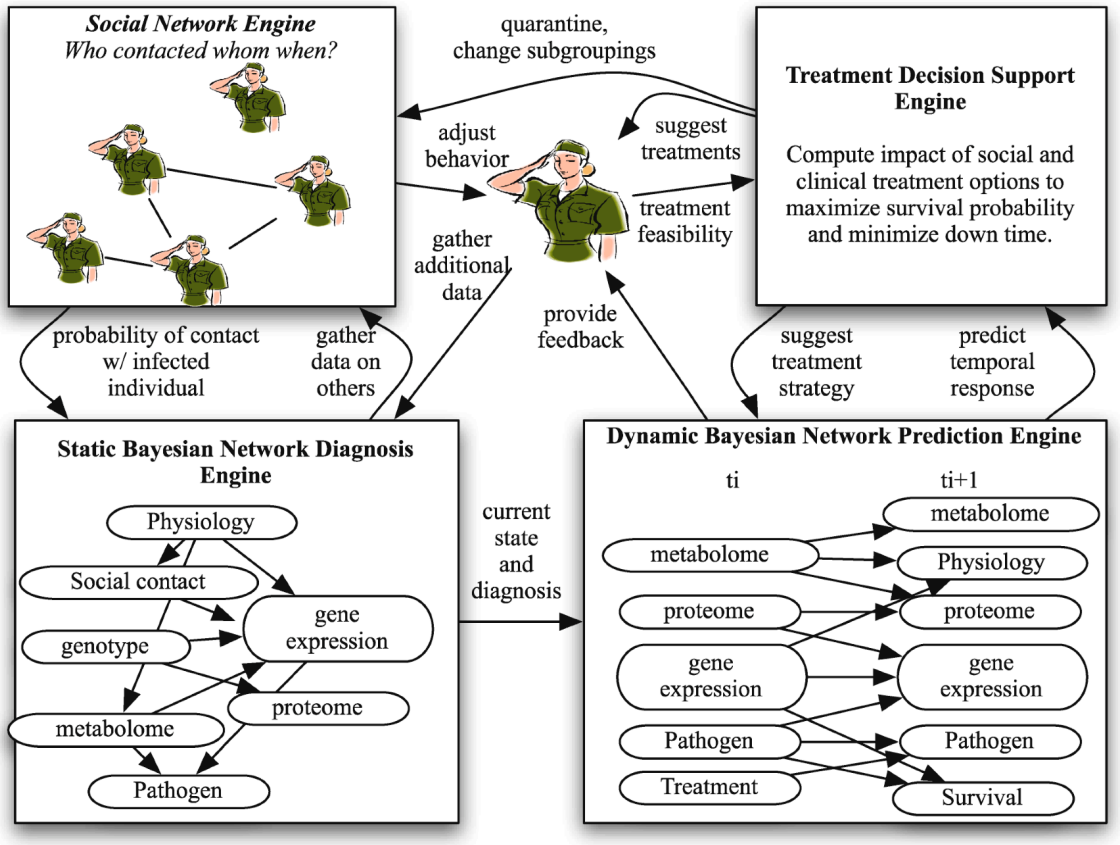


Figure 1.1: Overview of Predictive Health and Disease Socio-Molecular Inference Engine

Figure 1.1 decomposes into modules that are historically addressed in isolation from the others by a traditional field of research, such as epidemiology of bioinformatics. Epidemiologist have studied the propagation and evolution of infectious diseases across a population of individuals. Bioinformaticians aim at elucidating predictive patterns of high-dimensional molecular activity, e.g., gene expression, capable of discriminating between multiple phenotypes. Operations researchers in the public health domain aim at developing control and intervention developing strategies for partitioning individuals in a social network and strategically employing vaccinations to minimize the expected spread of disease. The ultimate goal of PHD, and the

four methods developed within this thesis, is to fuse these concepts into a singular decision support tool with an improvement in disease state detection over any one of these tools in isolation.

The methods presented in this dissertation are formulated as fundamental machine learning and statistical problems which are, in addition to PHD, applicable to a variety of problems in clinical medicine, quantitative finance, social-networking, statistical signal processing, and operations research.

1.2 Survey of Topics

The structure of this dissertation decomposes into four core chapters addressing a few fundamental methodological development necessary for a PHD inference engine.

Chapter 2, titled *Information Theoretic Adaptive Tracking in Complex Networks*, discusses the problem of sequentially identifying subsets of nodes in a complex network, e.g., social network, which elicit the largest expected gain in information regarding all observed and unobserved node states, e.g., phenotypes, in a network. Such resource allocation in the context of PHD would be the optimal distribution of few socio-molecular diagnostic devices to select individuals which will provide measurements which best minimize the expected information entropy of all hidden states in a large social network. The problem is formulated as a partially observed Markov decision process (POMDP) where the expected Kullback-Leibler (KL) divergence between a posterior distribution conditioned on a candidate set of measurements and a predicted posterior distribution devoid of any candidate measurements is used to select the best subset of nodes to sample. As updating the posterior distribution is a computationally difficult task, methods of approximate inference in Bayesian filtering are employed to tractably update the expected KL-divergence. Theoretical

analysis of the decay dynamics of the posterior posterior distribution under the *susceptible*, *infected*, *recovered* (SIR) graphical model of mathematical epidemiology are presented. An epidemic threshold is retrieved which depends on sensor likelihoods, the principle eigenvalue of the graph adjacency matrix, and the epidemiological parameters of transmission and recovery rates. Such a threshold indicates conditions when an epidemic is likely to affect the entire network. The chapter concludes with detection performance under the proposed adaptive sampling method on two synthetic complex networks.

The methodology in chapter 2 requires an estimate of the topology of the network of interest for the adaptive sampling algorithm. Chapter 3, titled *High Dimensional Spatio-Temporal Graphical Model Selection*, presents a tractable solution for estimating the structure of an SIR spatio-temporal graphical model using convex optimization when presented with historical observations of all node states in a network over time. Since the presence of an edge (connection) between two nodes is either present or absent, the problem of maximum likelihood estimation of the structure of the network is NP hard due to the combinatorial nature of enumerating through the space of possible networks. We propose relaxing the combinatorial variables in the log-likelihood distribution of an observed disease trajectory under the SIR model to a continuum where methods of convex optimization can be employed to tractably solve this problem. The likelihood of the observed disease trajectory is penalized with an ℓ_1-norm on the topological parameter vector which tends to promote a "sparse" estimate of the network structure, i.e., many edges are estimated as 0. Such spareness is a property of many complex networks, and ℓ_1-penalized likelihoods are commonly used to estimate the structure of sparse graphical models. The detection performance of the proposed method outperforms other state of the art discrete state graphical

model selection algorithms when detecting the topology of an SIR graphical model. The spatio-temporal grapihcal model selection procedure would allows a point-wise estimate of a network topology or a population of topologies (if a Bayesian viewpoint is desired), to be inserted into the adaptive sampling procedure detailed in chapter 2.

In the context of PHD, chapters 2 and 3 present solutions to the issue of reasoning under uncertainty about the paths and dynamics of transmission of an infectious agent on a social network and then how to optimally select individuals to sample. The latter two chapters involve discriminating between different phenotypes given the high-dimensional biomedical measurements resulting from sampling an individual. Specifically, chapter 4, titled *Robust Logistic Regression with Bounded Data Uncertainties*, extends the logistic regression classifier to be *robust* to bounded measurement error. As many biomedical measurements contain substantial measurement error, e.g., gene expression microarrays, one may wish to sacrifice optimality under "best-case" perturbations of the data and desire robustness to "worst-case" perturbations. Such robustness is desired in risk sensitive domains such as diagnosing patients. Block-sparsity promoting regularization penalties are added to loss function to accommodate the group sparse high-dimensional biomedical signals. The resulting thresholding conditions of both robustness and block-sparsity are presented and the relationship between group lasso regularization and group structured uncertainty is established. A block-coordinate gradient decent algorithm with iterative group thresholding is presented to solve this regularized robust logistic regression problem. In the limit of weakly separable data, the theoretical relationship between ridge logistic regression and robust logistic regression with spherical uncertainty is established. Under these asymptotic assumptions, the convergence rates of robust

vs. ridge are obtained and conditions when robust achieves convergence faster than ridge are extracted. The value added of using a robust logistic classifier is established by reporting smaller "worst-case" probability of error rates on gene expression data of patients inoculated with Human Rhino Virus (HRV) data set with ℓ_1-robust logistic regression over standard ℓ_1-logistic regression. In risk sensitive domains, such as clinical decision making, the improvement in "worst-case" detection performance of the proposed robust logistic regression classifier is desired in PHD when assigning a phenotype to a potential patient and any resulting actions that are conditioned on such labeling.

In PHD, and clinical medicine, one may be presented with a sequence of biomedical observations resulting from a series of patient visits to a clinic, thus producing a high-dimensional time-series. Often times the goal is discriminating between different phenotypes, e.g., symptomatic vs. asymptomatic, and one needs to identify the appropriate basis functions which summarize the high-dimensional, potentially mis-alligned, time-series. Chapter 5, titled *Functional Discriminants for Classification of High-Dimensional Time-Series*, models each time-series as a random function drawn from a phenotype dependent Gaussian Process (GPs). The log-odds ratio corresponding to each trajectory, e.g., gene observations over time, and each sample's time-series is then formed. The collection of these log-odds ratios from all variables and samples comprise the basis functions which are then used in forming a single linear classifier using ℓ_1-logistic regression. If the time-stamps corresponding to each observation in the time-series are unknown, one can place a prior distribution on these time-stamps and remove the dependence of time, thus providing additional flexibility of the resulting classifier in discriminating between phenotypes. The performance of this method is applied to a human data set involving patients inoculated

with Influenza A (H3N2), Respiratory syncytial virus (RSV), and HRV where the goal is discriminating between symptomatic (the patient developed symptoms after inoculation) or asymptomatic (the patient did not develop symptoms after inoculation). The ℓ_1-regularization paths exhibit the effect of "early predictive" genes vs. "late predictive" genes when trained on early or full subsets of the time-series upon perturbation with these viruses. The classification performance of the proposed method is presented using linear discriminant and quadratic discriminant GP basis functions and establish that one can discriminate between symptomatic and asymptomatic individuals prior to developing symptoms.

1.3 Contributions

The following journal publications, conference publications, and presentations represent the work detailed within this thesis.

Harrington Jr., P.L., and Hero III, A. O., Spatio-Temporal Graphical Model Selection (Submitted to *The Annals of Applied Statistics*). April 2010

Harrington Jr., P.L., Wiesel, A., and Hero III., A.O., Robust Logistic Regression with Bounded Data Uncertainty (In Preparation for the *IEEE Transactions on Signal Processing*). April 2010

Harrington Jr., P.L., Rao, A., and Hero III., A.O., Functional Discriminants for Classification and Prediction of Multiple Time-Series (In Preparation for the *Journal of Computational and Graphical Statistics*). April 2010

Harrington Jr., P.L., and Hero III, A. O., Information Theoretic Adaptive Tracking of Epidemics in Complex Networks *Forty-Seventh Annual Allerton Conference on Communication, Control, and Computing*, Monticello, IL USA, October 2009

Harrington Jr., P.L., and Hero III, A. O., Percolation Thresholds of Updated

Posteriors for Tracking Causal Markov Processes in Complex Networks. Technical Report. arXiv:0905.2236v1. March 2009

Harrington Jr., P.L., Rao, A., and Hero III, A. O., Classification and Subspace Selection of Multiple Biomedical Time-Series Data via Ensemble Learning. *AMIA Summit on Translational Bioinformatics*, San Francisco, CA USA, March 2009

Harrington Jr., P.L., Rao, A., and Hero III, A. O., Classification of Multiple Time-Series via Boosting. *IEEE Digital Signal Processing*, Marco Island, FL USA, January 2009

Harrington Jr., P.L., Rao, A., Kliger, M., Woolf, P.J., and Hero, A.O., Socio-Molecular Predictors of Health and Disease, *Invited Paper to 2008 Information Theory and Applications Meeting*, San Diego, CA. USA. January 2008

Harrington Jr., P.L., Rao, A., Kilger, M., Woolf, P.J., Hero, A.O., Spatio-Temporal Networks for Predictive Health and Disease. *IEEE Workshop on Statistical Signal Processing*, Madison, Wisconsin USA. August 2007

CHAPTER II

Information Theoretic Adaptive Tracking in Complex Networks

2.1 Introduction

This paper treats the important problem of monitoring the states of nodes in large computer, social, or power networks where these states dynamically change due to viruses, rumors, or failures that propagate according to the graph topology [9, 16, 40]. This class of network dynamics has been extensively modeled as a percolation phenomenon, where nodes on a graph can randomly "infect" their neighbors.

Percolation across networks has a rich history in the field of statistical physics, computer science, and mathematical epidemiology [27, 38, 40]. Here, researchers are typically confronted with a network, or a distribution over the network topology, and extract fixed point attractors of node configurations, thresholds for phase transitions in node states, or distributions of node state configurations [13, 6, 42]. In the field of fault detection, the nodes or edges can "fail", and the goal is to activate a subset of sensors in the network which yield high quality measurements that identify these failures [68, 51]. While the former field of research concerns itself with extracting *offline* statistics about properties of the percolation phenomenon on networks, devoid of any measurements, the latter field addresses *online* measurement selection tasks.

Here, we propose a methodology that actively tracks a causal Markov process

across a complex network (such as the one in Figure 2.3(a)), represented as a dynamic Bayesian network, where measurements are adaptively selected using feedback from the updated posterior distribution. We establish conditions such that the updated posterior probability of all nodes "infected" is driven to one as the number of time samples goes to infinity. The proposed epidemic/percolation threshold on the *updated* posterior distribution over the hidden states is a function of structural properties of the network, epidemiological parameters, and sensor likelihoods corresponding to those nodes that were sampled.

The proposed percolation threshold should more accurately reflect the true conditions that cause a phase transition in a network, e.g., node status changing from healthy/normal to infected/failed, than traditional thresholds derived from conditions on predictive distributions which are devoid of any measurements. As the conditions of a threshold are extracted by inspecting the dominant mode of decay of the updated posterior, this permits specification of best and worst case convergence rates of that a network clears the infection. Additionally, the decay dynamics of the updated posterior can yield insight into the asymptotic detection performance of the system for a given false alarm rate.

Since most practical networks of interest are large, it is usually infeasible to sample all nodes continuously and exhaustively. Given sampling constraints, we present an information theoretic sampling strategy that selects specific nodes that will yield the largest information gain, and thus, better detection performance.

The proposed sampling strategy balances the trade-off between trusting the *predictions* from the assumed model dynamics and expending precious resources to select a set of nodes for measurement.

We present the adaptive measurement selection problem and give two tractable

approximations to this subset selection problem based upon the joint and marginal posterior distribution, respectively. A set of decomposable Bayesian filtering equations are presented for this adaptive sampling framework and the tractable inference algorithms for complex networks are discussed. We present analytical worst case performance bounds for our adaptive sampling performance, which can serve as sampling heuristics for the activation of sensors or trusting predictions generated from previous measurements.

We believe that this is the first attempt to extract conditions of percolation thresholds in actively monitored dynamic Bayesian networks where the updated posterior distribution is the sufficient statistic of interest rather than observation independent predictive distributions.

2.2 Problem Formulation

The objective of actively monitoring the n node network is to recursively update the posterior distribution of each hidden node state given various measurements. Specifically, the next set of m measurement actions (nodes to sample), $m \ll p$, at next discrete time are chosen such that they yield the highest quality of *information* about the p hidden states. The condition on $m \ll p$ simulates the reality of fixed resource constraints, where typically only a small subset of nodes in a large network can be observed at any one time.

Here, the hidden states are discrete random variables that correspond to the states encoded by the percolation process on the graph. Here, the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V}$ representing the set of nodes and $\mathcal{E}$ corresponding to the set of edges. Formally, we will assume a state-space representation of a discrete time, finite state, partially

observed Markov decision process (POMDP). Here,

$$Z_k = \{Z_k^1, \ldots, Z_k^p\} \tag{2.1}$$

represents the joint hidden states, e.g., healthy or infected

$$Y_k = \{Y_k^{(1)}, \ldots, Y_k^{(m)}\} \tag{2.2}$$

represents the m observed measurements obtained at time k, e.g., biological assays or `PING`ing an IP address, and

$$a_k = \{a_k^1, \ldots, a_k^m\} \tag{2.3}$$

represents the m actions taken at time k, i.e., which nodes to sample. Here, $Y_k^{(j)}$, continuous/categorical valued vector of measurements, which is induced by action a_k^j, $a_k^j \in \mathcal{A}$, with $\mathcal{A} = \{1, \ldots, p\}$ confined to be the set of all p individuals in the graph, and $Z_k^i \in \{0, 1, \ldots, r\}$. Since the topology of $\mathcal{G}$ encodes the direction of "flow" for the process, the state equations may be modeled as a decomposable partially observed Markov process:

$$Y_k^i = f(Z_k^i) + w_k^i \tag{2.4}$$

$$Z_k^i = h(\{Z_{k-1}^j\}_{j \in \{\eta(i), i\}}). \tag{2.5}$$

Here, $\eta(i) = \{j : \mathcal{E}(\mathcal{V}_i, \mathcal{V}_j) \notin \emptyset\}$ is the neighborhood of i, $f(Z_k^i)$ is a non-random vector-valued function, w_k^i is measurement noise, and $h(\{Z_{k-1}^j\}_{j \in \{\eta(i), i\}})$ is a stochastic equation encoding the transition dynamics of the Markov process (see Figure 2.1 for a two node graphical model representation).

2.2.1 Bayesian Filtering

In our proposed framework for actively monitoring the hidden node states in the network, the posterior distribution is the sufficient statistic for inferring these states.

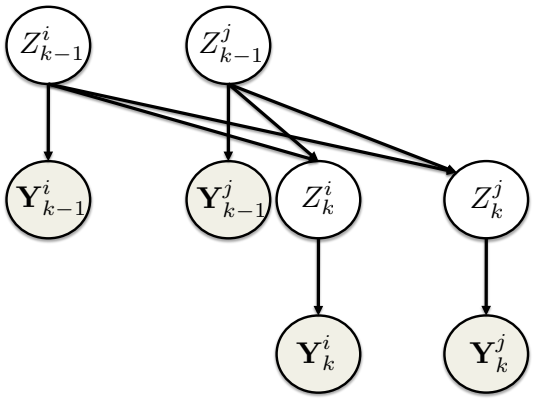


Figure 2.1: Partially Observed Markov Structure for i and j for $\mathcal{E}(\mathcal{V}_i, \mathcal{V}_j) \notin \emptyset$

The general recursion for updating the joint posterior probability given all past and present observations is given by the standard Bayes update formula:

$$p(Z_k|Y_{0:k}) = \frac{f(Y_k|Z_k)}{g(Y_k|Y_{0:k-1})} p(Z_k|Y_{0:k-1}) \tag{2.6}$$

with

$$p(Z_k|Y_{0:k-1}) = \sum_{z \in \{0,1,\ldots,r\}^p} p(Z_k|Z_{k-1} = z) p(Z_{k-1} = z|Y_{0:k-1}). \tag{2.7}$$

and

$$g(Y_k|Y_{0:k-1}) = \sum_{z \in \{0,1,\ldots,r\}^p} f(Y_k|Z_k = z) p(Z_k = z|Y_{0:k-1}). \tag{2.8}$$

The Chapman-Kolmogorov equations provide the connection between the posterior update (2.7) and the distribution resulting from the standard percolation equations. In the former, the updates are conditional probabilities that are conditional on past observations, while in the latter, the updates are not dependent on observations.

The local interactions in the graph $\mathcal{G}$ imply the following conditional independence assumptions:

$$f(Y_k|Z_k) = \prod_{i=1}^{p} f(Y^i_k|Z^i_k). \tag{2.9}$$

$$p(Z_k|Z_{k-1}) = \prod_{i=1}^{p} p(Z_k^i|\{Z_{k-1}^j\}_{j\in\{\eta(i),i\}}) \tag{2.10}$$

where the likelihood term is defined in (2.4) and the transition dynamics are defined in (2.5). This decomposable structure allows the belief state (posterior excluding time k observations) update, for the i^{th} node in $\mathcal{G}$, to be written as:

$$p(Z_k^i|Y_{0:k-1}) = \sum_{z\in\{0,1,\ldots,r\}^{\|pa(i)\|}} p(Z_k^i|Z_{k-1}^{pa(i)} = z)p(Z_{k-1}^{pa(i)} = z|Y_{0:k-1}) \tag{2.11}$$

with the parent set, $pa(i) = \{\eta(i), i\}$. Unfortunately, for highly connected nodes in $\mathcal{G}$, this marginal update becomes intractable. It thus must be approximated [12, 43, 35].

2.2.2 Information Theoretic Adaptive Sampling

In most real world situations, acquiring measurements from all p nodes at any time k is unrealistic, and thus, a sampling policy must be exploited for measuring a subset of nodes [24, 68, 8, 30]. Since we are concerned with monitoring the states of the nodes in the network, an appropriate reward is the expected information gain between the *updated* posterior, $p_k = p(Z_k|\{Y_k^i\}_{i\in a_k}, Y_{0:k-1})$, and the belief state, $p_{k|k-1} = p(Z_k|Y_{0:k-1})$:

$$a_k = \arg\max_{a\subset\mathcal{A}} \mathbb{E}\left[\mathcal{D}_\alpha\left(\{Y_k^i\}_{i\in a}\right)|Y_{0:k-1}\right] \tag{2.12}$$

$$\mathcal{D}_\alpha\left(\{Y_k^i\}_{i\in a}\right) = \mathcal{D}_\alpha\left(p_k||p_{k|k-1}\right),\ 0<\alpha<1 \tag{2.13}$$

with α-Divergence defined by

$$\mathcal{D}_\alpha(p||q) = \frac{1}{\alpha-1}\log\left(\mathbb{E}_q\left[(p/q)^\alpha\right]\right) \tag{2.14}$$

for any distributions p and q with identical support.

The reward in (2.12) has been widely applied to multi-target, multi-sensor tracking for many problems including, sensor management and surveillance [24, 47]. Note that

$\lim_{\alpha\to 1}\mathcal{D}_\alpha(p||q) \to \mathcal{D}_{KL}(p||q)$, where $\mathcal{D}_{KL}(p||q)$ is the Kullback-Leibler divergence between p and q. The expectation in (2.12) is taken with respect to the conditional distribution $g(Y_k|Y_{0:k-1})$ given the previous measurements $Y_{0:k-1}$ and actions a_k. In practice, the expected information divergence in (2.12) must be evaluated via Monte-Carlo methods. Also, the maximization in (2.12) requires enumeration over all $\binom{p}{m}$ actions (for subsets of size m), and therefore, we must resort to approximations. We propose incrementally constructing the set of actions at time k, a_k, for $j = 1, \ldots, m$, according to:

$$a_k^j = \text{argmax}_{i\in\mathcal{A}\setminus a_k}\mathbb{E}\left[\mathcal{D}_\alpha\left(Y_k^i, \{Y_k^j\}_{j\in a_k}\right)|Y_{0:k-1}\right]. \tag{2.15}$$

Both (2.12) and (2.15) are selecting the nodes to sample which yield maximal divergence between the percolation prediction distribution (belief state) and the updated posterior distribution, averaged over all possible observations. Thus (2.12) provides a metric to assess whether to trust the predictor and defer actions until a future time or choose to take action, sample a node, and update the posterior.

Lower Bound on Expected α-Divergence

Since the expected α-Divergence in (2.12) is not closed form, we could resort to numerical methods for estimating this quantity. Alternatively, one could specify an analytical lower-bound that could be used in-lieu of numerically computing the expected information gain in (2.12) or (2.15).

We begin by noting that the expected divergence between the updated posterior and the predictive distribution (conditioned on previous observations) differ only through the measurement update factor, $f_k/g_{k|k-1}$ ((2.12) re-written):

$$\mathbb{E}_{g_{k|k-1}}\left[\mathcal{D}_\alpha\left(p_k||p_{k|k-1}\right)\right]] = \mathbb{E}_{g_{k|k-1}}\left[\frac{1}{\alpha-1}\log\mathbb{E}_{p_{k|k-1}}\left[\left(\frac{f_k}{g_{k|k-1}}\right)^\alpha\right]\right] \tag{2.16}$$

where $f_k = f(Y_k|Z_k)$ and $g_{k|k-1} = g(Y_k|Y_{0:k-1})$. So, if there is significant overlap between the likelihood distributions of the observations, the expected divergence will tend to zero, implying that there is not much value-added in taking measurements, and thus, it is sufficient to use the predictive distribution for inferring the states.

It would be convenient to interchange the order of the conditional expectations in (2.16). It is easily seen that Jensen's inequality yields the following lower bound for the expected information gain

$$\mathbb{E}_{g_{k|k-1}}\left[\mathcal{D}_\alpha\left(p_k||p_{k|k-1}\right)\right] \geq \frac{1}{\alpha-1}\log \mathbb{E}_{p_{k|k-1}}\left[\mathbb{E}_{g_{k|k-1}}\left[\left(\frac{f_k}{g_{k|k-1}}\right)^\alpha\right]\right]. \tag{2.17}$$

Here, the inner conditional expectation can be obtained from $\mathcal{D}_\alpha\left(f_k||g_{k|k-1}\right)$, which has a closed form for common distributions (e.g., multivariate Gaussians) [24].

2.3 Asymptotic Analysis of Marginal Posterior

For tracking the percolation process across $\mathcal{G}$, we have discussed recursive updating of the posterior. However, computing these updates is generally intractable. For the remainder of the paper, we will use (2.4) and (2.5) to directly update the marginal posterior distribution using the following matrix representation:

$$p_k(z) = D_k(z)p_{k|k-1}(z) \tag{2.18}$$

with updated marginal posterior $p_k(z) = [p_{1,k}(z), \ldots, p_{p,k}(z)]^T$ with $p_{i,k}(z) = p(Z_k^i = z|Y_k^i, Y_{0:k-1})$, $D_k(z) = \text{diag}\left(f_{i,k}^{(z)}/g_{i,k|k-1}\right)$, and marginal belief state $p_{k|k-1}(z) = [p_{1,k|k-1}(z), \ldots, p_{p,k|k-1}(z)]^T$ with $p_{i,k|k-1}(z) = p(Z_k^i = z|Y_{0:k-1})$.

Note that for $i \notin a_k$, $(D_k(z))_{i,i} = 1$, and $p_{i,k}(z) = p_{i,k|k-1}(z)$. Given that we can find an efficient way of updating $p_{k|k-1}(z)$, according to the transition dynamics (2.5), we can solve a modified version of (2.15), for $j = 1, \ldots, m$:

$$a_k^j = \arg\max_{i\in\mathcal{A}\setminus a_k}\mathbb{E}\left[\mathcal{D}_\alpha\left(Y_k^i\right)|Y_{0:k-1}\right] \tag{2.19}$$

$$\mathcal{D}_\alpha\left(Y_k^i\right) = \mathcal{D}_\alpha\left(p_{i,k}(z)||p_{i,k|k-1}(z)\right),\ 0<\alpha<1. \tag{2.20}$$

2.3.1 Pearson χ^2 Divergence of Updated Marginal Posterior

One interesting property of the Bayesian filtering equations is that the updated posterior can be written as a perturbation of the predictive percolation distribution through the following relationship (z omitted for clarity):

$$p_k = D_k p_{k|k-1} = p_{k|k-1} + (D_k - I)\, p_{k|k-1}. \tag{2.21}$$

Hence, when the sensors do a poor job in discriminating the observations, $D_k \approx I$, we have $p_k \approx p_{k|k-1}$. It is of interest to determine when there is significant difference between the posterior update and the prior update specified by the standard percolation equations. Recall that the updated posterior is, in the mean, equal to the predictive distribution, $\mathbb{E}\left[p_k|Y_{0:k-1}\right] = p_{k|k-1}$. The total deviation of the updated posterior from the percolation distribution can be summarized by computing the trace of the following conditional covariance:

(2.22)

$$\mathrm{tr}\left(\mathrm{Cov}\left[p_k|Y_{0:k-1}\right]\right) = \mathrm{tr}\left(\mathbb{E}\left[\left(p_k - \mathbb{E}\left[p_k|Y_{0:k-1}\right]\right)\left(p_k - \mathbb{E}\left[p_k|Y_{0:k-1}\ \right]\right)^T|Y_{0:k-1}\right]\right).$$

Using (2.21) and properties of the trace operator, we obtain the following measure of total deviation of the updated posterior from the predictive distribution in terms of f_k and $g_{k|k-1}$:

$$\mathrm{tr}\left(\mathrm{Cov}\left[p_k|Y_{0:k-1}\right]\right) = \mathrm{tr}\left(\mathbb{E}\left[\left(D_k - I\right)^2|Y_{0:k-1}\right] P_{k|k-1}\right) \tag{2.23}$$

with $P_{k|k-1} = p_{k|k-1}p_{k|k-1}^T$. The conditional expectation in (2.23) is the Pearson χ^2 divergence between distributions $f_{i,k}$ and $g_{i,k|k-1}$, for all i. This joint measure of deviation is analytical for particular families of distributions and thus can be used as an alternative measure of divergence for activation of sensors [24].

2.3.2 Decay Dynamics of Updated Posterior Distribution

There has recently been significant interest in deriving the conditions of a percolation/epidemic threshold in terms of transition parameters and the graph adjacency matrix spectra for two state causal Markov processes [6, 13, 42]. Such thresholds yield conditions necessary for phase transition in the probability of local infections becoming epidemics. Knowledge of these conditions are particularly useful for designing "robust" networks, where the probability of epidemics is minimized.

Epidemic thresholds are typically obtained by extracting the sufficient conditions of the network and model parameters for the node states to be driven to their stationary point, with high probability. The probability of these events are computed using the observation independent distribution encoding the stochastic dynamics of the process [6, 13, 42].

We use the results in [6, 13] to derive a percolation threshold based upon the *updated* posterior distribution (2.6) assuming a restricted class of two-state Markov processes. The dominant mode of decay characterizing the conditions of a threshold should more accurately model the *current* dynamic response of the posterior distribution since the updated posterior tracks a particular "disease" trajectory better than the observation independent predictive distributions.

Formally, $Z_k^i \in \{0, 1\}$, $f_{i,k}^{(z)} = f(Y_k^i | Z_k^i = z)$ is the conditional likelihood for node i, $p_{i,k} = p(Z_k^i = 1 | Y_k^i, Y_{0:k-1})$, and $p_{i,k} = p(Z_k^i = 1 | Y_{0:k-1})$. Here, we will assume that $Z_k = 0$ is the unique absorbing state of the system.

The Bayes update for $p_{i,k}$ can be written as (i subscript omitted for clarity):

$$\begin{aligned} p_k &= \frac{f_k^{(1)}}{f_k^{(1)} p_{k|k-1} + f_k^{(0)}(1-p_{k|k-1})} p_{k|k-1} \\ &= \frac{f_k^{(1)}/f_k^{(0)}}{1+\frac{f_k^{(1)}-f_k^{(0)}}{f_k^{(0)}} p_{k|k-1}} p_{k|k-1} \\ &= \frac{f_k^{(1)}/f_k^{(0)}}{1+\frac{\Delta f_k}{f_k^{(0)}} p_{k|k-1}} p_{k|k-1}. \end{aligned} \tag{2.24}$$

There are three different sampling/observation dependent possibilities for each individual at time k: case (1), i is not sampled and therefore, $p_k = p_{k|k-1}$, case (2), $\Delta f_k > 0$, and case (3), $\Delta f_k < 0$. We first derive a tight-upper bound for cases (2) and (3) of the form $p_k \leq c_k \, p_{k|k-1}$. For the remainder of the analysis we will assume that $|\frac{\Delta f_k}{f_k^{(0)}} p_{k|k-1}| < 1$ for cases (2) and (3).

In case (2), when $\Delta f_k > 0$, we can re-write (2.24) in terms of an *alternating geometric series*:

$$\begin{aligned} p_k &= \frac{f_k^{(1)}}{f_k^{(0)}} \left[\sum_{l=0}^{\infty} (-1)^l \left(\frac{|\Delta f_k|}{f_k^{(0)}} p_{k|k-1} \right)^l \right] p_{k|k-1} \\ &\leq \frac{f_k^{(1)}}{f_k^{(0)}} \left[1 + \frac{|\Delta f_k|}{f_k^{(0)}} p_{k|k-1} \right] p_{k|k-1} \end{aligned} \tag{2.25}$$

where we have used the fact that $1/(1+|a|) \leq 1+|a|$. Recalling that $p \geq p^2$ for $0 \leq p \leq 1$, we have

$$p_k \leq \frac{f_k^{(1)}}{f_k^{(0)}} \left[1 + \frac{|\Delta f_k|}{f_k^{(0)}} \right] p_{k|k-1}. \tag{2.26}$$

In case (3), when $\Delta f_k < 0$, (2.24) can be expanded as a *geometric series*:

$$\begin{aligned} p_k &= \frac{f_k^{(1)}}{f_k^{(0)}} \left[\sum_{l=0}^{\infty} \left(\frac{|\Delta f_k|}{f_k^{(0)}} p_{k|k-1} \right)^l \right] p_{k|k-1} \\ &= \frac{f_k^{(1)}}{f_k^{(0)}} \left[1 + \frac{|\Delta f_k|}{f_k^{(0)}} p_{k|k-1} + \sum_{l=2}^{\infty} \left(\frac{|\Delta f_k|}{f_k^{(0)}} p_{k|k-1} \right)^l \right] p_{k|k-1}. \end{aligned} \tag{2.27}$$

Exploiting the fact that $p \geq p^2$ for $0 \leq p \leq 1$, we obtain:

$$p_k \leq \frac{f_k^{(1)}}{f_k^{(0)}} \left[1 + \frac{|\Delta f_k|}{f_k^{(0)}}\right] p_{k|k-1} + \mathcal{O}\left(\frac{|\Delta f_k|}{f_k^{(0)}} p_{k|k-1}\right) \tag{2.28}$$

where the higher order terms of $\frac{|\Delta f_k|}{f_k^{(0)}} p_{k|k-1}$ are captured in

$$\mathcal{O}\left(\frac{|\Delta f_k|}{f_k^{(0)}} p_{k|k-1}\right) = \sum_{l=2}^{\infty} \left(\frac{|\Delta f_k|}{f_k^{(0)}} p_{k|k-1}\right)^l. \tag{2.29}$$

Now that an upper-bound has been established for conditions when a node is sampled (under both scenarios on the signed difference of the sensor likelihoods), we can state the general equality/inequality (equality for case (1)) of $p_k \leq c_k \; p_{k|k-1}$ with

$$b_k = \begin{cases} 1 & , i \notin a_k \\ \frac{f_k^{(1)}}{f_k^{(0)}} \left[1 + \frac{|\Delta f_k|}{f_k^{(0)}}\right] & , |\Delta f_k| > 0 \end{cases}$$

with $c_k = b_k$ for cases (1) and (2) and $c_k = b_k + \mathcal{O}\left(\frac{|\Delta f_k|}{f_k^{(0)}} p_{k|k-1}\right)$ for case (3).

After gathering all p nodes into vector notation, we have the following element-wise upper-bound on the updated belief state:

$$p_k \leq C_k p_{k|k-1} = (B_k + \mathcal{O}_k)\, p_{k|k-1}. \tag{2.30}$$

with

$$B_k = \operatorname{diag}(b_{i,k}) \tag{2.31}$$

and

$$\mathcal{O}_k = \operatorname{diag}\left(\mathbb{I}_{\{\Delta f_{i,k}<0\}} \mathcal{O}\left(\frac{|\Delta f_{i,k}|}{f_{i,k}^{(0)}} p_{i,k|k-1}\right)\right) \tag{2.32}$$

where $\mathbb{I}_{\{\Delta f_{i,k}<0\}}$ is the indicator function for the event $\Delta f_{i,k} < 0$.

Thus far, we have established, under the assumptions of $|\frac{\Delta f_k}{f_k^{(0)}} p_{k|k-1}| < 1$, an upper-bound for the updated posterior in terms of observation likelihoods and the belief state (2.30).

Next, consider the restricted class of two-state Markov processes on $\mathcal{G}$, for which we can produce a bound of the form

$$p_{k|k-1} \leq S p_{k-1} \tag{2.33}$$

where S contains information about the transition parameters and the topology of the network.

A class of models where such a bound exist is the *Susceptible-Infected-Susceptible* (SIS) model of mathematical epidemiology [6]. The SIS model on a graph $\mathcal{G}$, assumes that each of the p individuals are in states 0 or 1, where 0 corresponds to *susceptible* and 1 corresponds to *infected*. At any time k, an individual can receive the infection from their neighbors, $\eta(i)$, based upon their states at $k-1$.

Under this SIS model, the matrix S is given by

$$S = (1-\gamma)I + \beta A \tag{2.34}$$

where the Markov transition parameters γ is the probability of i transitioning from 1 to 0, β is the probability of transmission between neighbors i and j, and A is the graph adjacency matrix (see Figure 2.2).

Returning to the derivation, using the bound on updating the belief state (2.33) and updating the posterior (2.30), we have by induction, the following recursion:

$$\begin{aligned} p_k &\leq C_k p_{k|k-1} \leq C_k S p_{k-1} \leq (C_k S \cdots C_1 S)\, p_0 \\ &= (B_k S \cdots B_1 S)\, p_0 + \mathcal{O}_{C_k S} \end{aligned} \tag{2.35}$$

where we have lumped the higher order modes and higher order cross-terms into $\mathcal{O}_{C_k S}$.

The *dominant mode of decay* of the updated posterior may be found by investi-

gating the following eigen-decomposition:

$$B_k S = \left(\sum_{j=1}^{p} b_{j,k} e_j e_j^T\right)\left(\sum_{j=1}^{p} \lambda_j u_j u_j^T\right) \tag{2.36}$$

with $e_j = [0, \ldots, 0, 1, 0, \ldots, 0]^T$ (1 at j^{th} element). Without loss of generality, we can assume the eigenvalues of S are listed in decreasing order, $|\lambda_1| \geq \cdots \geq |\lambda_p|$. Now rewriting (2.36), we have

$$\begin{aligned} B_k S &= \left(b_{j_k} e_{j_k} e_{j_k}^T + \mathcal{O}_B\right)\left(\lambda_1 u_1 u_1^T + \mathcal{O}_S\right) \\ &= \left(\lambda_1 b_{j_k} e_{j_k} e_{j_k}^T u_1 u_1^T + \mathcal{O}_{BS}\right) \end{aligned} \tag{2.37}$$

where $b_{j_k} = \max_{j\in\{1,\ldots,p\}} b_{j,k}$ and the $\mathcal{O}_B, \mathcal{O}_S, \mathcal{O}_{BS}$ variables corresponds to the higher order terms. Inserting (2.37) into (2.35), and matching the largest eigenvalues of B_k with λ_1 we obtain

$$\begin{aligned} p_k &\leq (B_k S \ldots B_1 S)\, p_0 + \mathcal{O}_{C_k S} \\ &= \lambda_1^k \prod_{l=1}^{k} b_{j_l} \left(\prod_{l=1}^{k} \left(e_{j_l} e_{j_l}^T u_1 u_1^T\right)\right) p_0 + \mathcal{O}(\varphi^k). \end{aligned} \tag{2.38}$$

Thus, at large k, the dominant mode of the posterior goes as $\lambda_1^k \prod_{l=1}^{k} b_{j_l}$ (the modes in $\mathcal{O}(\varphi^k)$ decay faster than the dominant mode presented above).

We can see that if the spectral radius of S is less than one, $|\lambda_1| < 1$, then for large k, $p_k \to 0$, which is the unique absorbing state of the system.

This epidemic threshold condition on λ_1 has been previously established for unforced (observation independent) SIS-percolation processes [6]. However, in the tracking framework, the rate at which the posterior decays to the *susceptible* state is damped by an additional measurement dependent factor, $\prod_{l=1}^{k} b_{j_l}$, resulting from using the *updated* posterior distribution.

This measurement-dependent dominant mode of the posterior should more accurately model the true dynamic response of the node states better than that in [6]

since the posterior better tracks the truth than the unforced predictive distribution. Additionally, this dominant mode of the updated posterior distribution allows one to simulate the response of the percolation threshold to intervention and control actions which are designed to increase the threshold, such that the probability of epidemics is minimized.

2.4 Numerical Example

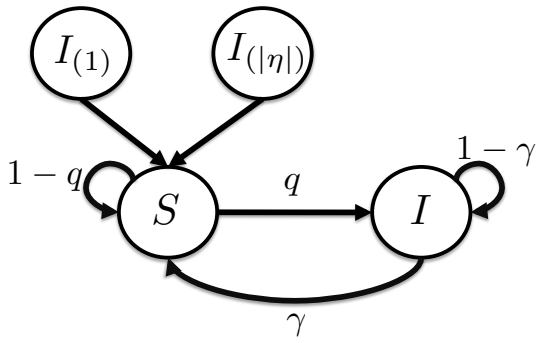


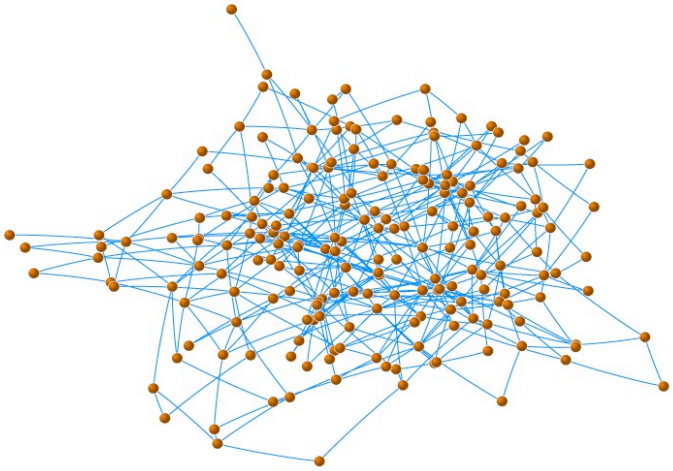
Figure 2.2: SIS Markov Chain for Node i Interacting with the Infected States of its Neighbors

Here, we present results of simulations of our adaptive sampling for the active tracking of a causal Markov ground truth process across a random 200 node, scale-free network (Figure 2.3(a)). Since most modern networks in which this method is most applicable, e.g., social networks, tend to be scale-free in their degree distribution (see Figure 2.3(b)), the proposed network shall suffice for extracting various statistics under the proposed adaptive tracking method. Since the goal in tracking is to accurately classify the states of each node, we are interested in exploring the detection performance as the likelihood of an epidemic increases through the percolation threshold for this network.

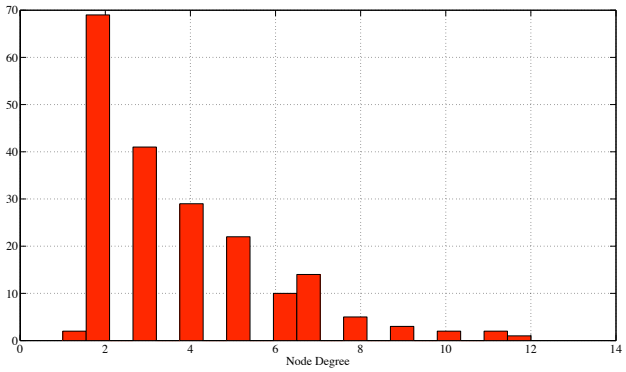
One would expect different phase transitions (thresholds) in detection performance for various sampling strategies, ranging from the lowest threshold for unforced predictive distributions to highest for a continuous monitoring of all n nodes. We will

present a few of these detection surfaces that depict these phase transitions for the unforced percolation distribution, random $m = 40$ node sampling, and our proposed information theoretic adaptive sampling of $m = 40$.

Here, we will restrict our simulations to the two-state SIS model of mathematical epidemiology described above.



(a) 200 Node Scale Free Synthetic Network



(b) Degree Distribution of the 200 Node Scale Free Network

Figure 2.3: Structure of the 200 Node Synthetic Network Used in Simulations

The sensor models (2.4), are of the form of two-dimensional multivariate Guassians with common covariance and shifted mean vector. The transition dynamics of the

i^{th} individual (2.5), for the SIS model is given by:

$$Z_k^i | Z_{k-1}^{\{i,\eta(i)\}} \sim (1-\gamma) Z_{k-1}^i + (1 - Z_{k-1}^i) \left[1 - \prod_{j \in \eta(i)} (1 - \beta Z_{k-1}^j) \right]. \tag{2.39}$$

where $Z_{k-1}^i \in \{0, 1\}$ is the indicator function of i being infected at time $k-1$. The transmission term between i and $\eta(i)$ is known the Reed-Frost model [6, 13, 40]. Since the tail of the degree distribution of our synthetic scale-free graph contains nodes with degree greater than 10, updating (2.11) exactly is unrealistic and we must resort to approximate algorithms. Here, we will assume the mean field approximation used by [6] for this SIS model, resulting in the following marginal belief state update for the i^{th} node of infected ($Z_k^i = 1$):

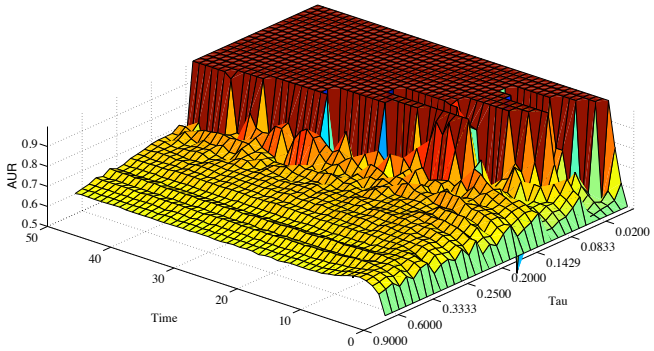
$$p_{i,k|k-1} = (1-\gamma) p_{i,k-1} + (1 - p_{i,k-1}) \left[1 - \prod_{j \in \eta} (1 - \beta p_{j,k-1}) \right]. \tag{2.40}$$

Equation (2.40) allows us to efficiently update the marginal belief state directly for all n nodes which are then used for estimating the best m measurements using (2.19).

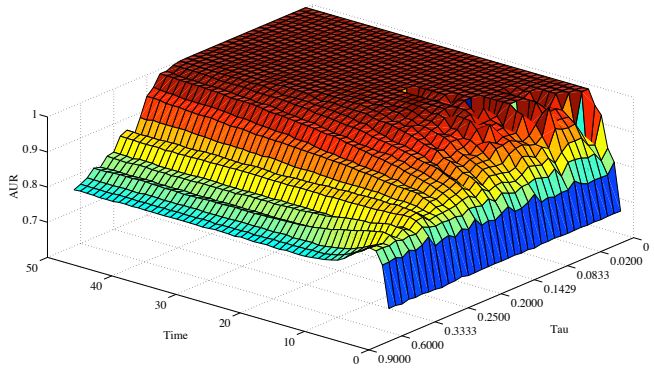
As we are interested in detection performance, as a function of time and epidemic intensity, the Area Under the ROC Curve (AUR) is a natural statistic to quantify the detection power at each time and each propensity of epidemic (detection of the infected state). The AUR is evaluated at each time k, each SIS percolation intensity parameter

$$\tau = \beta / \gamma \tag{2.41}$$

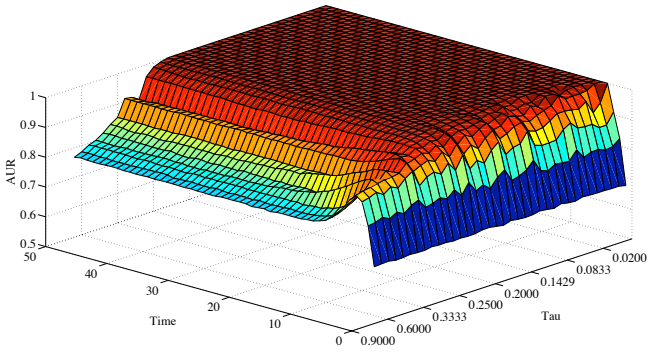
and over 500 random initial states of the network. For the SIS model, τ is the single parameter (aside from the topology of the graph) that characterizes the intensity of the percolation/epidemic. It is useful to understand how the detection performance varies as a function of epidemic intensity, as it indicates how well the updated posteriors are playing "catch-up" in tracking the true dynamics on the network.



(a) AUR Surface for Unforced Prediction Distribution (no evidence acquired throughout the monitoring)



(b) AUR Surface for Updated Posterior Distribution with $m = 40$ Random Measurements at Each Time k



(c) AUR Surface for Updated Posterior Distribution with $m = 40$ Information Theoretic Adaptive Measurements at Each Time k

Figure 2.4: Detection Performance Surface: Area Under the ROC (AUR) Curve Surface as a Function of Percolation Parameter $\tau = \beta/\gamma$ and Time

For this SIS model, the percolation threshold is defined as $\tau_c = 1/\lambda_1(A)$ where $\lambda_1(A) = \max_{i\in\{1,\ldots,p\}}|\lambda_i|$ is the spectral radius of the graph adjacency matrix, A [6].

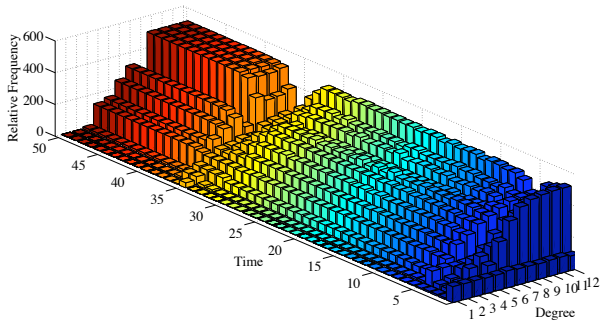
Values of τ greater than τ_c imply that any infection tend to become an epidemic, whereas those values less than τ_c imply that small epidemics tend to die out.

For the network under investigation (Figure 2.3(a)), $\tau_c = 0.1819$. We see from Figure 2.4(a) that a phase transition in detection power (AUR) for the unforced percolation distribution does indeed coincide with the epidemic threshold τ_c. While the epidemic threshold for the random and adaptive sampling policies is still $\tau_c = 0.1819$, the measurements acquired allow the posterior to better track the truth, but only up to their respective phase transitions in detection power (see Figures 2.4(b) and 2.4(c)).

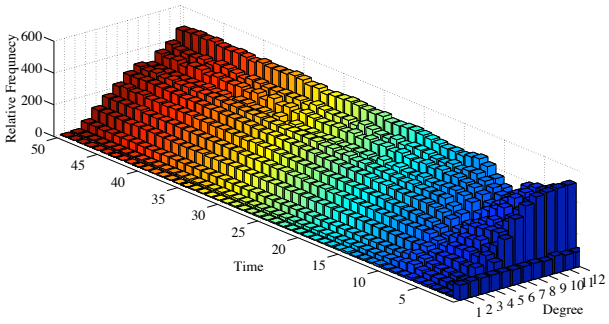
Figure 2.4(c) confirms that the adaptive sampling better tracks the truth than randomly sampling nodes, while pushing the phase transition in detection performance to higher percolation intensities, τ. We see that the major benefit of the adaptive sampling is apparent when conditions of the network are changing moderately, at medium epidemic conditions. Beyond a certain level of percolation intensity, more resources will need to be allocated to sampling to maintain a high level of detection performance.

A heuristic sampling strategy based on the topology of $\mathcal{G}$ was also explored (results not shown) by sampling the "hubs" (highly-connected nodes). However, detection performance was only slightly better than random sampling and poorer than our adaptive sampling method.

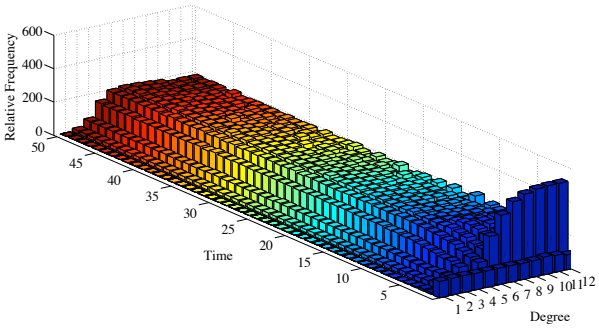
It is often useful for developing sampling heuristics and offline control/intervention policies to inspect what *type* of nodes, topologically speaking, is the adaptive sampling strategy targeting, under various network conditions (different values of τ). In Figure 2.5, the relative frequency of nodes sampled with a particular degree is plotted against time (under the $m = 40$ adaptive sampling strategy) for three different



(a) $\tau = 0.125 < \tau_c$



(b) $\tau = 0.2143 \approx \tau_c$



(c) $\tau = 0.5 < \tau_c$

Figure 2.5: Relative Frequency of Nodes Sampled of a Given Degree under $m = 40$ Adaptive Sampling Strategy

values of τ (over 500 random initial conditions of the network).

For the larger of the three values explored ($\tau = 0.5 > \tau_c$) we see that the sampling is approximately uniform across the nodes of each degree on the graph (Figure 2.5(c)). Therefore, under extremely intense epidemic conditions, the adaptive sampling strategy is targeting all nodes of each degree equally, and therefore, it is sufficient to perform random sampling. For the two lower values of τ, Figure 2.5(a) and Figure 2.5(b) (near τ_c), we see that adaptive policy targets highly connected

nodes more frequently than those of lesser degree and thus, it is more advantageous to exploit such a strategy, as compared to random sampling (see AUR surface in Figure 2.4(c)).

2.5 Conclusion

In this paper, we have presented an information theoretic framework for recursively selecting the best subset of nodes to sample in a dynamic Bayesian network that yield the largest expected information gain about the hidden state of the network. This framework can be applied to a variety of problem domains, including actively tracking an influenza outbreak across a population or adaptively monitoring the diffusion of information across large networks, such as a terrorism network.

Within the proposed adaptive tracking/sampling framework, we have derived conditions for a network specific percolation threshold using an updated posterior distribution rather than an observation independent predictive distribution. These conditions recover the unforced percolation threshold derived in [6] but with an additional factor involving sensor likelihood terms due to measurements obtained throughout the monitoring. A term of the form $\lambda_1^k \prod_{l=1}^k b_{j_l}$ (derived in (2.38)) was shown to be the dominant mode of the updated posterior dynamic response to active intervention of immunizing the nodes (holding node states constant). The conditions of the threshold, using the updated posterior, should more accurately model the phase transition in detection performance and thus, enable a better assessment of immunization strategies and any subsequent observations resulting from such actions. The framework and modes of decay of the updated posterior should provide additional insight into active monitoring of large complex networks under resource constraints.

Exploring phase transitions in updated posteriors for other classes of diffusion

is the subject of future work. One particularly interesting question is identifying conditions of phase transitions between multi-state processes (> 2) and explore the rates at which a system transitions between various states.

CHAPTER III

High Dimensional Spatio-Temporal Graphical Model Selection

3.1 Introduction

This chapter treats the problem of learning the interaction structure of a spatio-temporal graphical model for a discrete state and discrete time stochastic process known as the susceptible, infected, recovered (SIR) model. The presence of spatial interactions cause adjacent nodes in the graph to affect each others states over time. Learning the topology of this graph is known as model selection. We cast this graphical model selection problem as a penalized likelihood problem, resulting in a convex program for which convex optimization solvers can be applied. SIR spatio-temporal graphical models are commonly used in modeling the random propagation of information between nodes in large networks in bioinformatics, signal processing, public health, and national security [9, 16, 40]. Knowing the network link structure allows accurate prediction of individual node states and can aid the development of control and intervention strategies for such networks. This paper develops a tractable method to estimate the topology of the network for the SIR spatio-temporal graphical model from empirical data.

Exact solutions of the graphical model selection problem is NP hard due to the combinatorial nature of enumeration through the discrete space of possible graph

topologies. Researchers studying Bayesian networks, both static and dynamic, have developed exact and approximate methods for selecting a good candidate topology [10, 14, 39]. Such methods are appropriate for networks of small size and of unknown generative models for the observations. However, they are difficult to scale to larger graphs. SIR processes are used to model transmission events on complex networks which tend to be sparse in their interactions [41, 6, 40, 13, 9], so that there are relatively few edges in the graph. Over the past decade sparse regularization methods have been developed for graphical model selection using ℓ_1-regularization and other approaches. Examples include Gaussian graphical models (GMMs) [37, 18, 65, 50, 46] and Markov random fields (MRFs) [33, 59, 50].

The SIR model used throughout this paper is both discrete state and discrete time and thus any ℓ_1-penalized GMM method that is designed for real valued Gaussian random vectors would not be appropriate for this model. The structure learning algorithms for MRFs discussed in [33, 59, 50] are designed for discrete samples drawn from a MRF and most are limited to binary states (the SIR model has three states and is a different generative model than the MRF). Research in MRFs and GMMs have successfully used the ℓ_1-penalty to control the sparseness of the estimated graphical model topologies and we will adopt this approach for the SIR model. The method presented in this paper also scales to large networks more easily than traditional Bayesian network structure learning algorithms [10, 14, 39].

The proposed sparse structure learning method is designed for graphs that incorporate random causal transmission events affecting the evolution of the node states, such occurs in the propagation of infectious disease. Identifying the structure of social networks in tracking epidemics has received increased attention due to the global response to pandemic influenza A (H1N1) 2009. We illustrate the accuracy

of the proposed network structure learning on two moderate sized complex networks using real-world epidemiological parameters that approximate an H1N1 flu inspired outbreak [64]. We compare performance of the proposed estimation method against a MRF graphical model selection using ℓ_1-regularized logistic regression [59]. The proposed method is more accurate than generic approaches such as [59] for detection of anomalous network structure given sampled data from a spatio-temporal SIR process.

3.2 SIR Spatio-Temporal Graphical Models

The SIR graphical model has been used to approximate the general problem of modeling the evolution of node states in a network in which there is random transmission of disease or information between adjacent nodes on a graph [6, 40, 13, 9]. In the limit of large populations with equal mixing rates, SIR models have been used to model population proportions of particular states using differential equations [1, 48, 38, 27]. Unlike these studies, this paper addresses the problem of estimation of the topology of interactions between individual nodes in the network.

The SIR graphical model is a discrete time, discrete state model for the states of nodes in the network. Nodes can only affect the states of adjacent nodes in the network when they are in the "infected" state. The state of a node is given by $X_{i,k}$, where i refers to the individual (node) and k denotes time, and $X_{i,k}$ takes on values $x \in \{0, 1, 2\}$ (corresponding to "susceptible", "infected", and "recovered", respectively). The model is specified by the state transition probabilities given in

the 3x3 stochastic matrix

$$P_{i,k|k-1} = \begin{bmatrix} 1 - q_{i,k|k-1} & 0 & \gamma \\ q_{i,k|k-1} & 1-\alpha & 0 \\ 0 & \alpha & 1-\gamma \end{bmatrix} \tag{3.1}$$

where $q_{i,k|k-1}$ is the probability of transmission from "infected" neighbors of node i at time k, γ is the probability that node i transitions from "recovered" to "susceptible", and α is the probability that node i transitions from "infected" to "recovered". Since (3.1) allows a transition from recovered back to susceptible, this is actually a $SIRS$ model (SIR and $SIRS$ will be used interchangeably to refer to the three state stochastic process). For p nodes, the spatial topology of the network is defined by the interconnectivity, or adjacency, matrix

$$\mathcal{E} = \begin{bmatrix} \mathcal{E}_{1,1} & \cdots & \mathcal{E}_{1,p} \\ \vdots & \ddots & \vdots \\ \mathcal{E}_{p,1} & \cdots & \mathcal{E}_{p,p} \end{bmatrix} \tag{3.2}$$

where the l, m^{th} entry $\mathcal{E}_{l,m} \in \{0, 1\}$ is the indicator event that nodes l and m are connected. The pattern of non-zero entries in (3.2) specifies the interconnection topology of the network. The fundamental assumptions for an SIR network model is that the transition probabilities do not depend on node i while the interconnectivity matrix (3.2) is independent of time k. Under these assumptions, the joint distribution of an observed trajectory of length T, represented by the p-dimensional discrete state vector $X_k = [X_{1,k}, \ldots, X_{p,k}]^T$, factorizes

$$\mathbb{P}(X_1, \ldots, X_T) = \prod_{k=2}^{T} \mathbb{P}(X_k | X_{k-1}) = \prod_{k=2}^{T} \prod_{i=1}^{p} \mathbb{P}\left(X_{i,k} | \{X_{j,k-1}\}_{j \in \{\eta(i), i\}}\right), \tag{3.3}$$

where the neighborhood of node i is denoted

$$\eta_i = \{j : \mathcal{E}_{i,j} \neq 0\}. \tag{3.4}$$

The core component of most variations of the SIR model is the assumption that node i is conditionally independent of all non-neighboring nodes given the states of node i and its neighbors at time $k-1$. Each neighbor can transmit the "infection" to node i independent of the others neighbors. Under these assumptions, the probability of at least one transmission to node i at time k is given by

$$q_{i,k|k-1} = 1 - \prod_{j \in \eta_i} (1 - \omega z_{j,k-1}^{(1)}), \tag{3.5}$$

where $z_{k-1}^{(1)} \in \{0, 1\}$ is the indicator random variable of the j^{th} variable being in state "infected" at previous time $k-1$ and ω is the prior Bernoulli probability of transmission between j and i (also referred to as the attack rate). The conditional transition distribution in (3.3) is given by the following multinomial distribution

$$\mathbb{P}\left(X_{i,k} = x | \{X_{j,k-1} = x_{j,k-1}\}_{j \in \{\eta_i, i\}}\right) = \prod_{x \in \{0,1,2\}} \left(p_{i,k|k-1}(x)\right)^{z_{i,k}^{(x)}} \tag{3.6}$$

with indicator variable $z_{i,k}^{(x)} = \mathbb{I}_{\{x_{i,k}=x\}}$ and label probability $p_{i,k|k-1}(x)$ given by

$$p_{i,k|k-1}(x) = \begin{cases} \gamma z_{i,k-1}^{(2)} + z_{i,k-1}^{(0)} \prod_{j \in \eta_i} \left(1 - \omega z_{j,k-1}^{(1)}\right) & , x = 0 \\ z_{i,k-1}^{(0)} \left(1 - \prod_{j \in \eta_i} \left(1 - \omega z_{j,k-1}^{(1)}\right)\right) + (1 - \alpha) z_{i,k-1}^{(1)} & , x = 1 \\ \alpha z_{i,k-1}^{(1)} + (1 - \gamma) z_{i,k-1}^{(2)} & , x = 2, \end{cases}$$

where the model parameters are defined in (3.1). While the proposed graphical model selection method in this paper is motivated using the canonical three state SIR model, the method can be extended to any discrete state, discrete time stochastic model with state interactions of the form of the probability of transmission given in (3.5).

3.3 Spatio-Temporal Topology Estimation

Here we develop an estimate of the topology $\mathcal{E}$ (3.2) given training sequences $\mathcal{D}$ of observed states

$$\mathcal{D} = \{x_{i,k}\}_{i=1,k=1}^{p,T}, \tag{3.7}$$

where T is the horizon of the measurement period.

It will be convenient to rewrite the term involving the probability of transmission in (3.5) as

$$\begin{aligned}
\prod_{j\in\eta(i)} \left(1-\omega z_{j,k-1}^{(1)}\right) &= \exp\left\{\log\left(\prod_{j\in\eta_i}\left(1-\omega z_{j,k-1}^{(1)}\right)\right)\right\} \\
&= \exp\left\{\sum_{j\in\eta_i}\log\left(1-\omega z_{j,k-1}^{(1)}\right)\right\} \\
&= \exp\left\{\sum_{j\in\eta_i}\log\left(1-\omega\right) z_{j,k-1}^{(1)}\right\}, \qquad (3.8)
\end{aligned}$$

where we have exploited the fact that $\log(1-\omega z_{j,k-1}^{(1)}) = \log(1-\omega) z_{j,k-1}^{(1)} \leq 0$ in (3.8). Define $\theta_{i,j}$

$$\theta_{i,j} = \begin{cases} \log(1-\omega) & , \mathcal{E}_{i,j} = 1 \\ \\ 0 & , \mathcal{E}_{i,j} = 0 \end{cases}$$

and re-writing the sum term in (3.8) to run over the other $p-1$ nodes we arrive at the following

$$1 - q_{i,k|k-1} = \exp\left\{\sum_{j\neq i}\theta_{i,j} z_{j,k-1}^{(1)}\right\}, \ \theta_{i,j} \in \{\log(1-\omega), 0\} \ \forall j \neq i. \tag{3.9}$$

Inserting (3.9) into the state label probabilities, we have

$$p_{i,k|k-1}(x) = \begin{cases} \gamma z^{(2)}_{i,k-1} + z^{(0)}_{i,k-1} e^{\sum_{j\neq i}\theta_{i,j} z^{(1)}_{j,k-1}} & , x = 0 \\ z^{(0)}_{i,k-1}\left(1 - e^{\sum_{j\neq i}\theta_{i,j} z^{(1)}_{j,k-1}}\right) + (1-\alpha) z^{(1)}_{i,k-1} & , x = 1 \\ \alpha z^{(1)}_{i,k-1} + (1-\gamma) z^{(2)}_{i,k-1} & , x = 2. \end{cases}$$

Define the $p-1$ dimensional column vector θ_i by $\theta_i = \{\theta_{i,j}\}_{j\neq i}$. Given the spatial and temporal conditional independence assumptions represented in (3.3), the joint likelihood can be written as the multinomial distribution

$$\mathcal{L}(\phi; \mathcal{D}) = \prod_{k=2}^{T} \prod_{i=1}^{p} \prod_{x\in\{0,1,2\}} \left(p_{i,k|k-1}(x)\right)^{z^{(x)}_{i,k}} \tag{3.10}$$

with $\phi = \{\theta, \alpha, \gamma, \omega\}$ and $\theta = \{\theta_i\}_{i=1}^{p}$. The joint log-likelihood can be written as

$$\ell(\phi; \mathcal{D}) = \sum_{i=1}^{p} \ell(\phi_i; \mathcal{D}), \tag{3.11}$$

with $\phi_i = \{\theta_i, \alpha, \gamma, \omega\}$. The objective is to estimate the topology parameter θ while the α, γ, and ω are nuisance parameters. The i^{th} log-likelihood function is

$$\begin{aligned} \ell(\theta_i; \mathcal{D}) &= \sum_{k=2}^{T} \left\{ z^{(0)}_{i,k} \log p_{i,k|k-1}(0) + z^{(1)}_{i,k} \log(p_{i,k|k-1}(1)) \right\} \\ &= \sum_{k=2}^{T} \left\{ z^{(0,0)}_{i,k|k-1} \sum_{j\neq i} \theta_{i,j} z^{(1)}_{j,k-1} + z^{(1,0)}_{i,k|k-1} \log\left(1 - e^{\sum_{j\neq i}\theta_{i,j} z^{(1)}_{j,k-1}}\right) \right\}, \end{aligned} \tag{3.12}$$

with $z^{(0,0)}_{i,k|k-1} = z^{(0)}_{i,k} z^{(0)}_{i,k-1}$ and $z^{(1,0)}_{i,k|k-1} = z^{(1)}_{i,k} z^{(0)}_{i,k-1}$. Note that (3.12) only includes the state transition probabilities that involve $\theta_{i,j}$ since $\theta_{i,j}$ is obtained by optimizing over $\ell(\theta_i; \mathcal{D})$. In particular, the transition from any state to recovered does not depend on $\theta_{i,j}$. Note that the only parameter appearing in (3.12) necessary for estimation of θ is the transmission attack rate ω, appearing implicitly through the definition of $\theta_{i,j}$, $\theta_{i,j} \in \{\log(1-\omega), 0\}$.

Maximization of the likelihood over all possible $\theta \in \{\log(1-\omega), 0\}^{p(p-1)}$ is intractable even for small networks. The key to our maximum likelihood estimation approach is to relax $\theta_{i,j}$ to a continuous valued variable lying between its discrete values $\log(1-\omega)$ and 0, i.e., $\log(1-\omega) \leq \theta_{i,j} \leq 0$.

We use an ℓ_1-penalty on the likelihood to enforce sparsity, i.e., only a few $\theta_{i,j}$ are non-zero. Such ℓ_1-penalization is common in high dimensional statistical problems [55, 59, 29, 37, 18, 65, 50]. This yields the following convex program

$$\begin{aligned} &\min_\theta -\ell(\theta; \mathcal{D}) + \lambda \|\theta\|_{\ell_1} \\ &\text{s.t. } \log(1-\omega) \preceq \theta \preceq 0 \end{aligned} \tag{3.13}$$

with $\lambda > 0$ and $\preceq$ denotes element wise inequality between vectors. The estimated neighborhood set of node i is then

$$\hat{\eta}_i(\lambda) = \{j : \hat{\theta}_{i,j}(\lambda) < 0\}. \tag{3.14}$$

The set of all such neighborhoods will specify a (directed) graph that can be used to estimate the network topology $\mathcal{E}$ in (3.2). Specifically, the estimate of the l^{th} m^{th} entry of $\mathcal{E}$ by $\hat{\mathcal{E}}_{l,m}(\lambda) = \mathbb{I}_{\{\hat{\theta}_{l,m}(\lambda)<0\}}$. The global estimate of the topology is then defined as $\hat{\mathcal{E}}(\lambda) = \{\hat{\mathcal{E}}_{l,m}(\lambda)\}_{l,m}$.

3.3.1 Incorporating Prior Knowledge

There generally exists prior topological constraints that couple the optimization over $\{\theta_i\}_{i=1}^p$ for different i in (3.11). One such topological constraint is symmetry in the interactions, i.e., $\theta_{i,j} = \theta_{j,i}$, corresponding to an undirected graph $\mathcal{E}$. One way to incorporate this symmetry is to use augmented lagrangian methods that impose symmetry in the form of a variational penalty, e.g., $\sum_{i,j}(\theta_{i,j} - \theta_{j,i})^2$ [44]. Another method is to relax the symmetry constraint during the optimization followed by averaging the $\theta_{i,j}$ and $\theta_{i,j}$ together after optimization is completed.

If symmetry in $\theta_{i,j}$ is not imposed, the joint log-likelihood naturally factorizes as in (3.11), and can be decoupled by applying a coordinate descent-like likelihood function maximization that cycles through different nodes, updating its neighborhoods and holding the other θ_i's fixed:

$$\min_{\theta_i} -\ell(\theta_i; \mathcal{D}) + \lambda \sum_{j \neq i} |\theta_{i,j}|$$

$$\text{subject to } \log(1-\omega) \leq \theta_{i,j} \leq 0, \ \forall j \neq i. \tag{3.15}$$

Researchers may have additional prior knowledge such as known interactions, known non-interactions, or minimum or maximum size of neighborhoods. Some common forms of prior knowledge, and their corresponding constraints are summarized in Table 3.1.

Table 3.1: Common prior knowledge for complex networks appearing as constraints for the SIR graphical model selection problem

Prior Knowledge	Form of Constraint
Symmetry	$\theta_{i,j} = \theta_{j,i}$
Known Interactions	$\theta_{i,j} = \log(1-\omega), j \in \eta_i$
Known Non-Interactions	$\theta_{i,j} = 0, j \notin \eta_i$
Min Possible Size of Neighborhood	$\sum_{j \neq i} \theta_{i,j} \geq a_i \cdot \log(1-\omega)$
Max Possible Size of Neighborhood	$\sum_{j \neq i} \theta_{i,j} \leq b_i \cdot \log(1-\omega)$

It is more natural to work with the dual of the objective function in (3.13). In the dual one can immediately identify which of the inequality constraints are active. For instance, if one has prior knowledge regarding the maximum size of a particular

neighborhood, e.g., $\sum_{j \neq i} \theta_{i,j} \leq b \cdot \log(1-\omega)$, one can determine if $b \cdot \log(1-\omega) < s$, in which case, the constraint of $\|\theta\|_{\ell_1} \leq s$ would be inactive for the subvector θ_i. This results in convexity preserving topological constraints

$$
\begin{aligned}
(3.16) \qquad & \min_{\theta} -\ell(\theta; \mathcal{D}) \\
& \text{subject to} \quad \|\theta\|_{\ell_1} \leq s \\
& \log(1-\omega) \preceq \theta \preceq 0 \\
& \{h_j(\theta) \leq \nu_j\}_{j=1}^{k} \\
& \{g_l(\theta) = 0\}_{l=1}^{r}.
\end{aligned}
$$

3.3.2 Numerical Solution

The proposed ℓ_1-penalized likelihood problem in (3.15) is a convex program where there exists a variety of powerful solvers capable of producing a solution [4]. The proposed numerical solution in this paper is most appropriate for networks on the order of hundreds to a few thousand nodes. For networks on the order of tens of thousands of nodes, a large scale method such as the one given in [29] might be more appropriate.

We will relax the symmetry constraints when optimizing over θ and later impose them as a post-estimation heuristic

$$
(3.17) \qquad \hat{\eta}_i(\lambda^*) \leftarrow \hat{\eta}_i(\lambda^*) \cup j, \text{if } i \in \hat{\eta}_j(\lambda^*) \cap j \notin \hat{\eta}_i(\lambda^*) \forall i, j.
$$

We use a coordinate-wise gradient descent based method for solving (3.15) by quadratically expanding the negative log-likelihood, resulting in iteratively solving a sequence of quadratic programs that incorporates an additional line search. The Newton-step

update is accomplished by solving

$$\delta\theta_i^{(m)} = \arg\min_{\theta_i} \frac{1}{2}\theta_i^T H_i^{(m)}\theta_i + \theta_i^T g_i^{(m)} + \lambda \sum_{j\neq i} |\theta_{i,j}|$$

$$\text{s.t.} \quad \log(1-\omega) \leq \theta_{i,j} \leq 0, \ \forall j \neq i, \tag{3.18}$$

with gradient

$$g_i^{(m)} = -\nabla \ell(\theta_i; \mathcal{D})\big|_{\theta_i = \hat{\theta}_i^{(m)}}, \tag{3.19}$$

and Hessian

$$H_i^{(m)} = -\nabla^2 \ell(\theta_i; \mathcal{D})\big|_{\theta_i = \hat{\theta}_i^{(m)}}. \tag{3.20}$$

The updated parameter $\hat{\theta}_i^{(m+1)}$ given by

$$\hat{\theta}_i^{(m+1)} = \hat{\theta}_i^{(m)} + \epsilon_i^{(m)} \delta\theta_i^{(m)}, \tag{3.21}$$

with step size $\epsilon_i^{(m)}$ determined by performing a backtracking line search [4]

(3.22)

$$\text{while } -\ell(\hat{\theta}_i^{(m)} + \epsilon_i^{(m)}\delta\theta_i^{(m)}; \mathcal{D}) > -\ell(\hat{\theta}_i^{(m)}; \mathcal{D}) + 0.2\epsilon_i^{(m)}(g_i^{(m)})^T \delta\theta_i^{(m)}, \epsilon_i^{(m)} \leftarrow 0.3\epsilon_i^{(m)},$$

with $\epsilon_i^{(m)}$ initially set to 1. While (5.35) is convex, the presence of the ℓ_1-norm makes the objective function non-differentiable. However, the objective function can be transformed into an equivalent convex, differentiable objective by replacing the ℓ_1-norm with linear inequality constraints [4, 29]. An alternative to solving the Newton update (5.35) with the $(p-1)$x$(p-1)$ Hessian is replace it with a quasi Newton update which construct a surrogate objective function [57, 36, 32] and replaces the Hessian, $H_i^{(m)}$ with $\alpha_i^{(m)} I$, where I is the identity and $\alpha_i^{(m)}$ is chosen such that

$$\alpha_i^{(m)} I \succeq H_i^{(m)}, \tag{3.23}$$

and (3.23) means that $\alpha_i^{(m)} I - H_i^{(m)} \succeq 0$ is positive semi-definite. A consequence of the proposed penalized likelihood formulation for the SIR model is that $H_i^{(m)}$, in addition to being symmetric and positive semi-definite, has positive entries, i.e., $(H_i^{(m)})_{s,r} \geq 0$. Such non-negative conditions on the entries in $H_i^{(m)}$ can be enforced by using the Perron-Frobenius bound [25]

$$\max_s \lambda_s \left(H_i^{(m)} \right) \leq \max_s \sum_r \left(H_i^{(m)} \right)_{s,r}, \tag{3.24}$$

where the optimization is applied to the upper bound

$$\alpha_i^{(m)} = \max_s \sum_r \left(H_i^{(m)} \right)_{s,r}, \tag{3.25}$$

thus guaranteeing (3.23).

By replacing the Hessian with a diagonal surrogate is that the $p-1$-dimensional quadratic program in (5.35) factorizes into $p-1$ individual programs which have an analytical update and can be evaluated simultaneously. The update for $\theta_{i,j}$ under such a surrogate Hessian becomes

$$\begin{aligned} \delta\theta_{i,j}^{(m)} &= \arg\min_{\theta_{i,j}} \frac{1}{2}\alpha_i^{(m)}\theta_{i,j}^2 + g_{i,j}^{(m)}\theta_{i,j} + \lambda|\theta_{i,j}|,\ \log(1-\omega) \leq \theta_{i,j} \leq 0 \\ &= \begin{cases} \frac{-1}{\alpha_i^{(m)}} \left(|g_{i,j}^{(m)}| - \lambda \right)_+ & : g_{i,j}^{(m)} < \lambda - \alpha_i^{(m)} \cdot \log(1-\omega) \\ \log(1-\omega) & : g_{i,j}^{(m)} \geq \lambda - \alpha_i^{(m)} \cdot \log(1-\omega) \end{cases} \end{aligned} \tag{3.26}$$

with $(u)_+ = \max(0, u)$. The proposed gradient descent method for the ℓ_1-penalized likelihood problem for the spatio-graphical model selection problem is summarized in Algorithm 1 (below).

Algorithm 1

1. Let $\hat{\theta}_i^{(0)} \in [\log(1-\omega), 0]^{p-1}$ be an initial parameter vector for the i^{th} neighborhood.

2. Update $\delta\theta_i^{(m)}$ by solving (5.35) or solving (3.26) $\forall j \neq i$ with surrogate diagonal Hessian given by (3.25)

3. $\epsilon_i^{(m)} \leftarrow$ backtracking line search from (3.22)

4. $\hat{\theta}_i^{(m+1)} = \hat{\theta}_i^{(m)} + \epsilon_i^{(m)} \delta\theta_i^{(m)}$

5. If convergence criteria met, stop and repeat step 1 with next node index, $i \leftarrow i+1$. If convergence criteria not met, update gradient and Hessian (and potentially the surrogate diagonal Hessian) and repeat step 2 through 5 Note: Algorithm 1 can be parallelized across all p log-likelihoods rather than the cyclical update of $i \leftarrow i+1$. Symmetry is imposed through (3.17).

A possible speed up would be to perform active set updates to those coefficients which are non-zero by preferentially updating the coefficients corresponding to nodes that most likely belong to the neighborhood. Such active set updates have been used successfully in estimating sparse partial correlations [46]. They have also been proposed to block co-ordinate descent in group lasso logistic regression [36]. Implementing such accelerations is out of the scope of this paper.

3.3.3 Selection of Tuning Parameters

Algorithm 1 requires specification of the tuning parameter λ. Typically, an estimate of the best λ is desirable in order to perform cross validation or other error assessment. In this paper we report a BIC-like penalty, similarly used in previous work on the estimation of partial correlation networks [46], for selecting the best estimate of λ, denoted by λ^*, by cross validation. Specifically, assuming the attack rate ω is known, we perform the update $\theta_{i,j}$ as follows

$$\hat{\theta}_{i,j}(\lambda) \leftarrow \log(1-\omega), \forall i,j \in \{i,j : \hat{\theta}_{i,j}(\lambda) < 0\}. \tag{3.27}$$

The BIC penalty for the i^{th} node is

$$BIC_i(\lambda) = -\ell_i(\hat{\theta}_i(\lambda); \mathcal{D}) + \frac{1}{2}\log T_i \ \#\{j : \hat{\theta}_{i,j}(\lambda) < 0\}, \tag{3.28}$$

where $\#\{i, j : \hat{\theta}_{i,j}(\lambda) < 0\}$ is the number of non-zero entries in the estimator. The term $T_i = \#\{k : z_{i,k}^{(0)} = 1\}$ represents the effective time horizon for the i^{th} node as the number of terms in the i^{th} log-likelihood, which depends on the number of of $z_{i,k}^{(0)}$ equal to one (see (3.12)). Given (3.28), there will be multiple regularization parameters, one for each neighborhood i:

$$\lambda_i^* = \arg\min_\lambda BIC_i(\lambda). \tag{3.29}$$

The common approach is to impose that all the λ_i^*'s are the same and solve for a single tuning parameter

$$\lambda^* = \arg\min_\lambda \sum_{i=1}^{p} BIC_i(\lambda). \tag{3.30}$$

The latter approach has been previously used in controlling the sparseness of estimated partial correlation networks [46] and learning directed acyclic graphs (DAGs) [53].

3.4 Model Selection Under Observation Errors

In practice, the training trajectories are often contaminated through observation mislabeling. There are two common ways for accommodating for measurement error: (1) - maximum likelihood over the hidden true node states that are randomly perturbed through observation or (2) - robust estimation of the hidden true nodes states. Within the negative maximum likelihood framework, the former approach reduces to solving a *min min* program, which is non-convex, where 2^{p-1} hidden combinatorial state variables need to be estimated under Bernoulli measurement noise corruption.

In risk sensitive domains, one might sacrifice optimality under best-case mislabeling errors and instead, desire an estimate of a neighborhood structure that is robust to *worst-case* mislabeling. This robust approach [15, 2] produces an analytically tractable robust negative log-likelihood which is then minimized with respect to θ_i. Directly extracting the joint robust log-likelihood is intractable. We will approximate this joint robust likelihood by identifying the robust marginal log-likelihood corresponding to worst-case mislabeling of all other nodes necessary for estimating the i^{th} neighborhood structure. For simplicity of presentation, we will confine our attention to the two-state SIS process, however, the results within this section can be naturally extended to any multi-state variant of the SIR model.

The robust graphical model selection problem is given via the following

$$\begin{aligned} &\min_{\theta_i} \max_{y\in\mathcal{Z}(z,m)} -\ell(\theta_i;\mathcal{D}(y)) + \lambda\sum_{j\neq i}|\theta_{i,j}| \\ &\text{subject to } \log(1-\omega)\leq\theta_{i,j}\leq 0,\ \forall j\neq i \end{aligned} \tag{3.31}$$

where the uncertainty set is given by

$$\mathcal{Z}(z,m) = \{\mathcal{Z}_k(z,m)\}_{k=2}^{T} \tag{3.32}$$

with uncertainty set at time k given by

(3.33)

$$\mathcal{Z}_k(z,m) = \left\{z_{j,k-1} : y_{j,k-1} = z_{j,k-1} + \delta_{j,k-1}(1-2z_{j,k-1}), j\neq i, \delta_{k-1}\in\{0,1\}^{p-1}, \mathbf{1}^T\delta_{k-1}\leq m\right\}$$

where $z_{j,k-1} = z^{(1)}_{j,k-1}$ and $y_{j,k-1} = y^{(1)}_{j,k-1}$ have had their superscripts removed for clarity. When $\delta_{j,k-1} = 1$, a mislabeling occurs, and we are alloted at most m per time k. Under these assumptions, the marginal negative log-likelihood for node i can

be written as

(3.34)
$$-\ell(\theta_i;\mathcal{D}(y)) = -\sum_{k=2}^{T}\left\{(1-z_{i,k})\theta_i^T\psi_{k-1}(y) + z_{i,k}\,\log\left(1-e^{z_{i,k-1}\log\gamma+\theta_i^T\psi_{k-1}(y)}\right)\right\}$$

with $\psi_{k-1}(y) = \{(1-z_{i,k-1})y_{j,k-1}\}_{j\neq i}$.

As the inner maximization in (3.31) only affects the negative log-likelihood, we proceed with evaluating the maximization step via the following

$$\begin{aligned}
\Phi &= \max_{y\in\mathcal{Z}(z,m)} -\ell(\theta_i;\mathcal{D}(y)) \\
&= \max_{y\in\mathcal{Z}(z,m)} -\sum_{k=2}^{T}\left\{(1-z_{i,k})\theta_i^T\psi_{k-1}(y) + z_{i,k}\,\log\left(1-e^{z_{i,k-1}\log\gamma+\theta_i^T\psi_{k-1}(y)}\right)\right\} \\
&= \max_{\delta\in\Delta} -\sum_{k=2}^{T}\left\{(1-z_{i,k})\theta_i^T\psi_{k-1} + (1-z_{i,k})\Omega(\delta_{k-1}) + z_{i,k}\,\log\left(1-e^{z_{i,k-1}\log\gamma+\theta_i^T\psi_{k-1}+\Omega(\delta_{k-1})}\right)\right. \\
&= \max_{\delta\in\Delta} -\sum_{k=2}^{T}\left\{(1-z_{i,k})\Omega(\delta_{k-1}) + z_{i,k}\,\log\left(1-e^{z_{i,k-1}\log\gamma+\theta_i^T\psi_{k-1}+\Omega(\delta_{k-1})}\right)\right\}
\end{aligned}$$

where uncertainty set Δ is given by

(3.35)
$$\Delta = \left\{k : \delta_{k-1}\in\{0,1\}^{p-1}, \mathbf{1}^T\delta_{k-1}\leq m\right\}$$

and $\Omega(\delta_{k-1})$ is given by

(3.36)
$$\Omega(\delta_{k-1}) = (1-z_{i,k-1})\theta_i^T\left(I-2D_{k-1}\right)\delta_{k-1}$$

with $D_{k-1} = \mathrm{diag}(z_{k-1})$.

The estimate of δ_{k-1}, denoted by δ_{k-1}^* is conditional on three possible outcomes of the node states $\{z_{i,k}=1, z_{i,k-1}=0\}$, $\{z_{i,k}=0, z_{i,k-1}=0\}$, and $z_{i,k-1}=1$.

When $z_{i,k-1}=1$ the influence of θ_i is removed from the likelihood and the ability to resolve interactions is destroyed and thus, our attention will be devoted to the other two conditions.

Under the conditions when $z_{i,k} = 1$ and $z_{i,k-1} = 0$, the inner maximization of (3.31) reduces to solving the following

$$\begin{aligned} \max_{\delta_{k-1}\in\Delta} -\Omega(\delta_{k-1}) &= \max_{\delta_{k-1}\in\Delta} -\theta_i^T \left(I - 2D_{k-1}\right)\delta_{k-1} \\ &= \max_{\delta_{k-1}\in\Delta} -\sum_{j\neq i} \theta_{i,j}\left(1 - 2z_{j,k-1}\right)\delta_{j,k-1} \\ &= \max_{\delta_{k-1}\in\Delta} \sum_{j:z_{j,k-1}=0} \tilde{\theta}_{i,j}\delta_{j,k-1} - \sum_{j:z_{j,k-1}=1} \tilde{\theta}_{i,j}\delta_{j,k-1} \end{aligned} \tag{3.37}$$

with $\tilde{\theta}_{i,j} = -\theta_{i,j} \geq 0$. In (3.37), the first sum is positive whereas the second sum is negative, thus our attention is devoted to the first term where $z_{j,k-1} = 0$. Here, if the number of variables in the 0 (susceptible) state is less than or equal to m, e.g., $|\{j : z^{(1)}_{j,k-1} = 0\}| \leq m$ then the estimation is given by

$$\delta^*_{k-1} = \{1\}_{j\in\mathcal{I}^{(0)}_{k-1}} \cup \quad \{0\}_{j\notin\mathcal{I}^{(0)}_{k-1}}. \tag{3.38}$$

with $\mathcal{I}^{(0)}_{k-1} = \{j \neq i : z_{j,k-1} = 0\}$.

In the case when more variables are in the zero-state than the allotted m labeling errors, $|\{j : z_{j,k-1} = 0\}| > m$, we relax the combinatorial uncertainty set (3.35) to the following

$$\Delta = \left\{k : \delta_{k-1} \in [0,1]^{p-1}, \|\delta_{k-1}\|_{\ell_2} \leq \sqrt{m}\right\}. \tag{3.39}$$

Isolating the terms in the first sum in (3.37) and setting $\delta_{j,k-1} = 0$ for all $z_{j,k-1} = 1$ (as they only decrease the value of the objective function) we have

$$\begin{aligned} \sum_{j:z_{j,k-1}=0} \tilde{\theta}_{i,j}\delta_{j,k-1} &= \left[(I - D_{k-1})\tilde{\theta}_i\right]^T \delta_{k-1} \\ &= \left(I_{k-1}\tilde{\theta}_i\right)^T \delta_{k-1} \\ &\leq \|\theta_i\|_{I_{k-1}} \|\delta_{k-1}\|_{\ell_2} \\ &\leq \sqrt{m}\|\theta_i\|_{I_{k-1}}. \end{aligned} \tag{3.40}$$

The upper-bound in (3.40) is achieved at the max of (3.37) and therefore

$$\max_{\delta_{k-1}\in\Delta} - \Omega(\delta_{k-1}) = \sqrt{m}\|\theta_i\|_{I_{k-1}}. \tag{3.41}$$

Both (3.37) (with worst-case mis-labeling indexed by (3.38)) and (3.45) are the solutions to the k^{th} inner-maximization step of (3.31) for $\{z_{i,k} = 1, z_{i,k-1} = 0\}$ with $|\{j : z_{j,k-1} = 0\}| \leq m$ and $|\{j : z_{j,k-1} = 0\}| > m$, respectively. We will now proceed to solve the k^{th} inner-maximization step of (3.31) when $\{z_{i,k} = 0, z_{i,k-1} = 0\}$.

Conditioning on $\{z_{i,k} = 0, z_{i,k-1} = 0\}$, the maximization of the k^{th} component of the negative log-likelihood in (3.31) reduces to the following

$$\max_{\delta_{k-1}\in\Delta} -\log\left(1 - e^{\theta_i^T\psi_{k-1}+\Omega(\delta_{k-1})}\right) = -\log\left(1 - e^{\theta_i^T\psi_{k-1}+\max_{\delta_{k-1}\in\Delta}\Omega(\delta_{k-1})}\right). \tag{3.42}$$

Inspecting the maximization over (3.36) within the argument of the negative logarithm in (3.42) we have

$$\begin{aligned}
\max_{\delta_{k-1}\in\Delta}\Omega(\delta_{k-1}) &= \max_{\delta_{k-1}\in\Delta} \theta_i^T \left(I - 2D_{k-1}\right)\delta_{k-1} \\
&= \max_{\delta_{k-1}\in\Delta} \sum_{j\neq i} \theta_{i,j}\left(1 - 2z_{j,k-1}\right)\delta_{j,k-1} \\
&= \max_{\delta_{k-1}\in\Delta} \sum_{j:z_{j,k-1}=1} \tilde{\theta}_{i,j}\delta_{j,k-1} - \sum_{j:z_{j,k-1}=0} \tilde{\theta}_{i,j}\delta_{j,k-1}. \qquad (3.43)
\end{aligned}$$

If $|\{j : z_{j,k-1} = 1\}| \leq m$, then each of the corresponding $\delta^*_{j,k-1} = 1$ and the rest are set to zero, $\mathcal{I}_{k-1} = \{j : z_{j,k-1} = 1\}$. However, if $|\{j : z_{j,k-1} = 1\}| > m$ then we must relax the uncertainty set to (3.39) and proceed similarly as before. Isolating the terms in the first sum of (3.43) and setting $\delta_{j,k-1} = 0$ for all those in the second sum (as they only decrease the value of the objective function) we have the following

$$\begin{aligned}
\sum_{j:z_{j,k-1}=1} \tilde{\theta}_{i,j}\delta_{j,k-1} &= \left[D_{k-1}\tilde{\theta}_i\right]^T \delta_{k-1} \\
&\leq \|\theta_i\|_{D_{k-1}}\|\delta_{k-1}\|_{\ell_2} \\
&\leq \sqrt{m}\|\theta_i\|_{D_{k-1}}. \qquad (3.44)
\end{aligned}$$

The upper-bound in (3.44) is achieved at the max of (3.43) and therefore

$$\max_{\delta_{k-1}\in\Delta}\Omega(\delta_{k-1}) = \sqrt{m}\|\theta_i\|_{D_{k-1}}. \tag{3.45}$$

Given the mis-labelings under worst-case conditions, the "robust" penalized likelihood problem is

$$\begin{aligned} &\min_{\theta_i} \; -\ell_r(\theta_i;\mathcal{D}) + \lambda\sum_{j\neq i}|\theta_{i,j}| \\ &\text{subject to } \; \log(1-\omega) \leq \theta_{i,j} \leq 0, \; \forall j \neq i \end{aligned} \tag{3.46}$$

with robust negative log-likelihood given by

$$-\ell_r(\theta_i;\mathcal{D}) = -\sum_{k:z_{i,k-1}=0}\left\{(1-z_{i,k})\Omega^*_{k-1} + z_{i,k}\,\log\left(1-e^{\theta_i^T\psi_{k-1}+\Omega^*_{k-1}}\right)\right\} \tag{3.47}$$

and values of Ω^*_{k-1} presented in Table 3.2.

Table 3.2: Conditional values of Ω^*_{k-1} for robust marginal likelihood

Ω^*_{k-1}	Conditions
$\theta_i^T\,(I-2D_{k-1})\,\delta^*_{k-1}$	$\{z_{i,k}=1, z_{i,k-1}=0\}$ and $\lvert\{j : z_{j,k-1}=0\}\rvert \leq m$
$-\sqrt{m}\|\theta_i\|_{I_{k-1}}$	$\{z_{i,k}=1, z_{i,k-1}=0\}$ and $\lvert\{j : z_{j,k-1}=0\}\rvert > m$
$\theta_i^T\,(I-2D_{k-1})\,\delta^*_{k-1}$	$\{z_{i,k}=0, z_{i,k-1}=0\}$ and $\lvert\{j : z_{j,k-1}=1\}\rvert \leq m$
$\sqrt{m}\|\theta_i\|_{D_{k-1}}$	$\{z_{i,k}=0, z_{i,k-1}=0\}$ and $\lvert\{j : z_{j,k-1}=1\}\rvert > m$

The non-zero entries of δ^*_{k-1} in rows 1 and 3 are given by $\mathcal{I}_{k-1} = \{j : z_{j,k-1}=0\}$ and $\mathcal{I}_{k-1} = \{j : z_{j,k-1}=1\}$, respectively.

3.5 Numerical Results

Given the global response to the recent outbreak of pandemic influenza A (H1N1) 2009, the ability of public health organizations and world governments to develop

effective control and intervention strategies depends on knowledge of the topology of social networks. We illustrate the proposed penalized likelihood topology estimate for the problem of identifying the structure of synthetic social networks given disease spread that has attack rate parameters that simulate H1N1, specifically $\omega = 0.273$ as reported in [64]. The other two parameters, not needed for network inference but necessary for generating SIR trajectories from (3.2), were taken as $\alpha = 0.250$ reflecting a mean infectious period of 4 days and $\gamma = 0.100$ producing an average time of 10 days for transition from "recovered" to "susceptible".

We simulated two 200 node networks using two types of connection models: scale-free and small-world. These models have been proposed for many practical complex networks [41]. The two randomly generated networks used for experiments were created using the iGraph package for R [11]. The power law network was sampled such that the degree distribution reflected those which appear in real complex networks. Specifically, the exponent parameter of the degree distribution was taken as 2.2, consistent with evidence reported in [41]. The rewiring probability of the small-world network was taken as 0.1 to elicit tight communities that were loosely connected to other clusters.

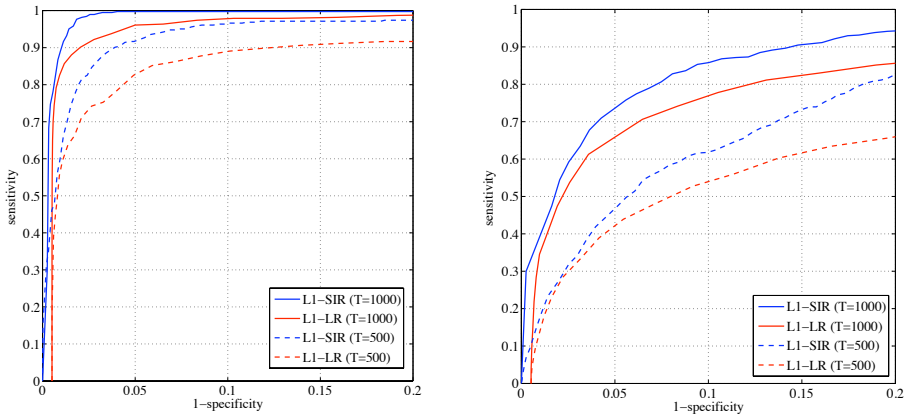
The SIR model (3.2) was used to generate training, validation, and test data for each of the two simulated 200 node networks. The networks were initialized with 40 randomly selected nodes were in "infected" state while the rest were in "susceptible" state. The quadratic program appearing in Algorithm 1, (5.35), was solved using the CVX environment in MATLAB and the solver SDPT3 4.0 [21, 58] with cold start initializations of $\hat{\theta}_i^{(0)} = 0$. Symmetry was included in the estimated neighborhoods following the post estimation heuristic (3.17).

We present a comparison against a modified version of graphical model selec-

tion using ℓ_1-logistic regression (ℓ_1-LR) [59]. Since the method described in [59] is designed for binary random variables generated from an Ising model, to implement ℓ_1-LR we transform the three state SIR variables to binary random variables. The transformation is the following: for each node i, the indicator event of the i^{th} node transitioning from "susceptible" to "infected", is regressed on all other $p-1$ "infected" nodes indicator variables at previous time $k-1$ with a bias controlling constant as explained in [59] and symmetry imposed through (3.17). By transforming the multi-state SIR random variables to the binary random variables for the implementation of ℓ_1-LR, we capture the causality of transmission from neighbors. While we transform the three state SIR random variables to two state random variables for implementing ℓ_1-LR, the proposed graphical model selection in this paper, referred to as ℓ_1-SIR, uses the original three state variables in the log-likelihood (3.12). As the estimated parameters using ℓ_1-LR [59] can take on any value on the real-line, we define the estimated neighborhood for the i^{th} node as those estimates with non-zero value.

The ROC curves corresponding to ℓ_1-SIR and the modified ℓ_1-LR for the scale-free network and small-world network for $T = \{500, 1000\}$ are displayed in Figure 3.1(a) and Figure 3.1(b), respectively. Inspection of Figure 3.1 validates that the proposed ℓ_1-SIR graphical model selection outperforms ℓ_1-LR when confronted with data drawn from the SIR distribution. At a false alarm rate of 5%, we see that the proposed ℓ_1-SIR method achieves a $5\% - 10\%$ gain in power over the modified ℓ_1-LR method for both networks. As ℓ_1-LR [59] uses one ℓ_1 penalty, for baseline comparison between these two graphical model selection algorithms, only a single regularization penalty was used in the ROC curves generated from ℓ_1-SIR. Both structure learning methods perform poorer in the case of the small-world network than in the case of

the scale-free network. This is possibly due to the increased frequency of re-infection in the tight clusters of the small-world network.

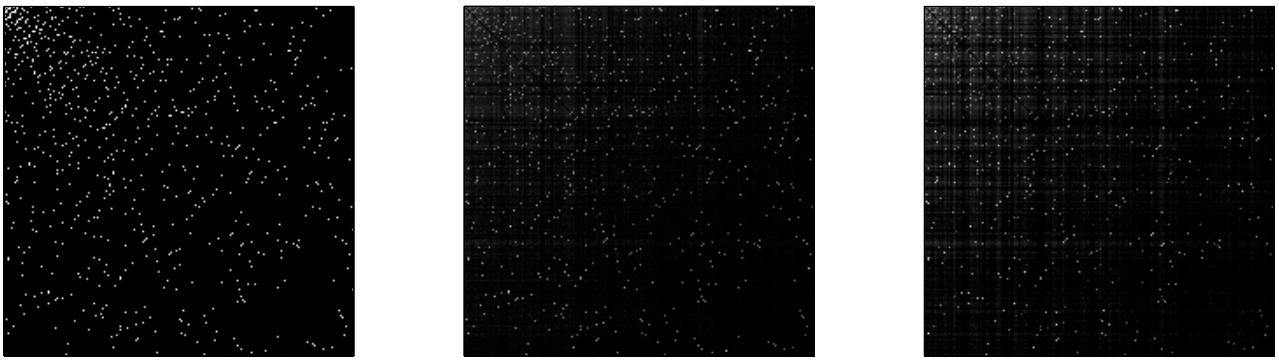


(a) Scale-Free Network (b) Small-World Network

Figure 3.1: ROC curves of ℓ_1-SIR graphical model selection (blue) vs. ℓ_1-logistic regression (red) for number of time points $T = \{500, 1000\}$

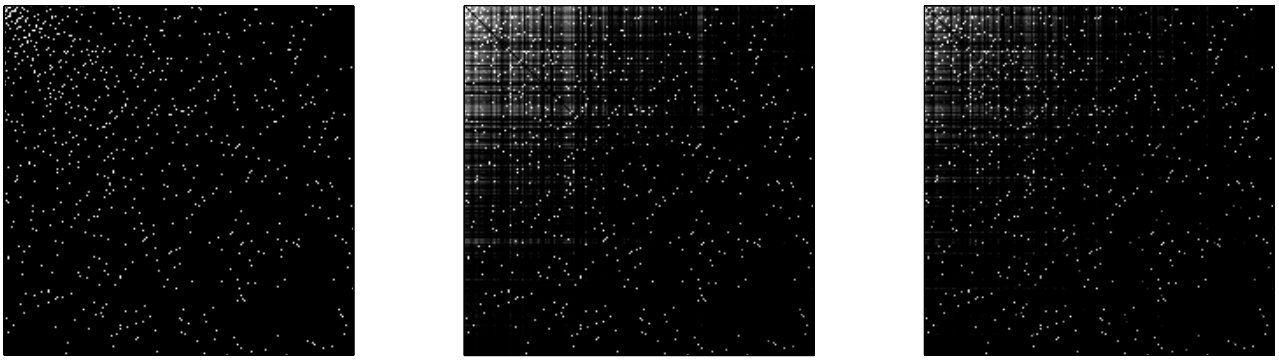
We next present the model selection performance on the 200 node scale-free network using the proposed method with global and neighborhood specific penalties, optimized by minimizing the BIC penalties (3.30) and (3.29), respectively, for time durations of $T = \{100, 400, 700, 1000\}$. The images in Figures 3.2 and 3.3 reflect the estimated network topologies, represented as symmetric adjacency matrices $\mathcal{E}$, averaged over the 1000 resampled initial conditions corresponding $T = 100$ and $T = 1000$, respectively. Subfigures a.) through c.) correspond to ground truth, ℓ_1-SIR with a single ℓ_1-penalty, and ℓ_1-SIR with neighborhood specific ℓ_1-penalty, respectively.

The intensity, located at row i and column j, indicates the frequency of an edge discovered between nodes i and j, white designates a strong edge and black designates no edge. Visual inspection of these figures establish that the proposed ℓ_1-SIR graphical model selection methods accurately extract the global community structure of the scale-free network when using a single or multiple penalties to enforce sparseness.



a.) b.) c.)

Figure 3.2: % zeros in the reconstruction of edges in 200 node synthetic scale free network under 100 time points resampled over 1000 initial conditions of 40 randomly selected nodes as "infected" with rest "susceptible". a.) ground truth, b.) single tuning parameter, c.) multiple tuning parameters (white - 0% black 100%)



a.) b.) c.)

Figure 3.3: % zeros in the reconstruction of edges in 200 node synthetic scale free network under 1000 time points resampled over 1000 initial conditions of 40 randomly selected nodes as infected with rest susceptible. a.) ground truth, b.) single tuning parameter, c.) multiple tuning parameters (white - 0% black 100%)

A quantitative comparison of accuracy of topology estimation is given by the sensitivity, specificity and probability of error. Table 3.3 summarizes the mean (with standard deviation shown in parentheses) when assessing the performance across the 1000 reconstructed topologies corresponding to the 1000 resampled simulations. We see that the sensitivity of this method, using a single λ^* and multiple $\{\lambda_i^*\}_{i=1}^p$, increases when the number of time samples increases while the specificity remains robust to the number of time samples and consistently above 0.96. Likewise, the global probability of error is below 0.05 for both methods across all time horizons explored. It is worth noting that the proposed method is only able to resolve an interaction between nodes i and j if both nodes states have changed at some point throughout the monitoring interval. Therefore for small time horizons, the epidemic may not have enough time to propagate the entire graph thus inhibiting the ability to accurately detect interactions.

A scale-free network has a wide distribution of vertex degrees (few hubs, many lesser connected nodes). Figure 3.4 a.), b.), and c.) show the sensitivity, specificity, and probability of error, respectively, of correctly detecting the neighborhood of each node as a function of increasing vertex degree. In all three subfigures, we see that regularizing with tuning parameters characteristic to each neighborhood $\{\lambda_i^*\}_{i=1}^p$ selected according to (3.29) tends to produce similar sensitivity and specificity with lower probability of error across all types of node degrees than when regularizing with a single penalty λ^* selected according to (3.30).

The performance of the proposed method was also assessed for a 200 node small-world network. Visual inspection of Figure 3.6 shows that the proposed method method accurately extracts the small-world community structure, represented by

Table 3.3: Detection statistics vs. time horizon for 200 node synthetic scale-free network with trajectories resampled over 1000 initial conditions of 40 randomly selected nodes as "infected" with rest "susceptible"

Method	T	$Sens.(\lambda^*)$	$Spec.(\lambda^*)$	$P_e(\lambda^*)$
ℓ_1-SIR(λ^*)	100	0.40(0.02)	0.96(0.00)	0.05(0.00)
ℓ_1-SIR($\{\lambda_i^*\}_{i=1}^p$)	100	0.34(0.02)	0.97(0.00)	0.05(0.00)
ℓ_1-SIR(λ^*)	400	0.80(0.05)	0.97(0.00)	0.03(0.00)
ℓ_1-SIR($\{\lambda_i^*\}_{i=1}^p$)	400	0.78(0.05)	0.96(0.00)	0.04(0.00)
ℓ_1-SIR(λ^*)	700	0.95(0.08)	0.96(0.00)	0.04(0.00)
ℓ_1-SIR($\{\lambda_i^*\}_{i=1}^p$)	700	0.95(0.07)	0.96(0.00)	0.04(0.00)
ℓ_1-SIR(λ^*)	1000	0.97(0.08)	0.96(0.00)	0.03(0.00)
ℓ_1-SIR($\{\lambda_i^*\}_{i=1}^p$)	1000	0.97(0.08)	0.96(0.00)	0.03(0.00)

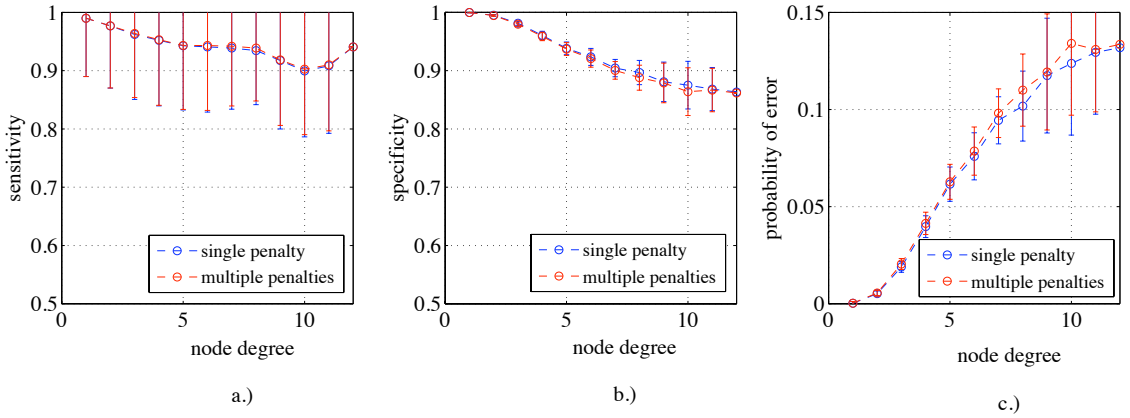
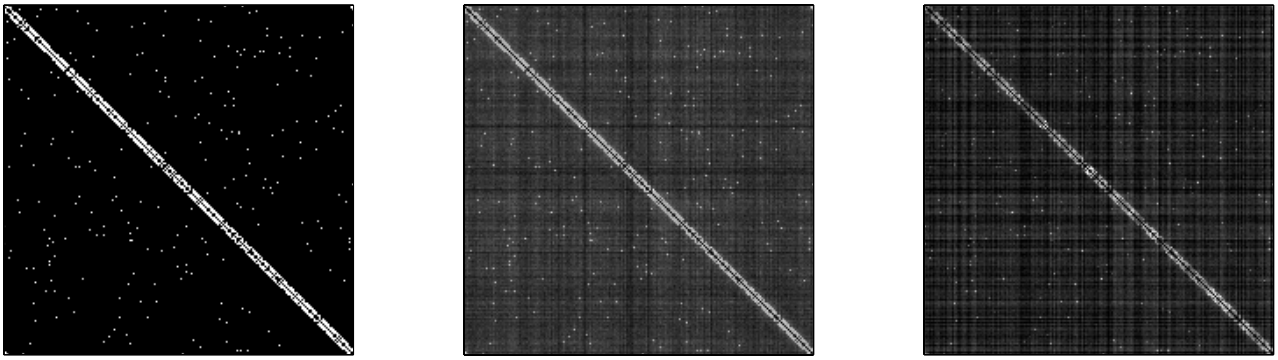


Figure 3.4: Neighborhood detection statistics vs. node degree for 200 node scale-free network with $T = 1000$ with trajectories resampled over 1000 initial conditions of 40 randomly selected nodes as infected with rest susceptible. a.) sensitivity, b.) specificity, c.) probability of error (red Single Penalty, blue Multiple Penalties)

the recovery of the banded structure of the adjacency matrices. In addition to detecting the characteristic clusters of the small-world ground truth network, the method also tends to identify the between-cluster interactions which are depicted in the off-diagonal elements. In terms of the detection statistics (Table 3.4), the sensitivity of both methods improves with the number of time samples and the single



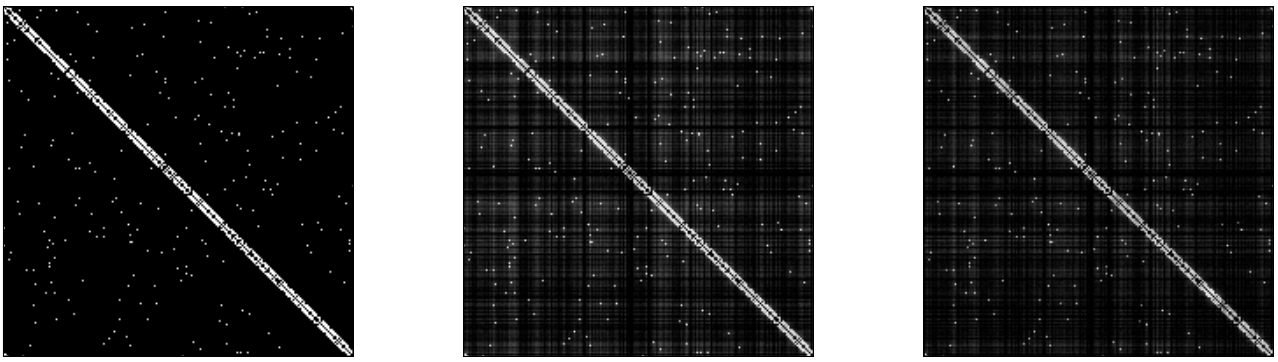
a.) b.) c.)

Figure 3.5: % zeros in the reconstruction of edges 200 node synthetic small world network under 100 time points resampled over 1000 initial conditions of 40 randomly selected nodes as "infected" with rest "susceptible". a.) ground truth, b.) single tuning parameter, c.) multiple tuning parameters (white - 0% black 100%)

tuning parameter method (3.30) results in higher power across all time samples. The method of regularizing with tuning parameters unique to each neighborhood (3.29) seems to perform similarly to the method when using a single penalty. The decomposition of the global detection statistics on a per vertex degree basis for the small-world network was also explored. Figure 3.7 a.), b.), and c.) represent the sensitivity, specificity, and probability of error, respectively, in reconstructing the neighborhoods of nodes as a function of node degree. The more highly connected nodes tend to have poorer sensitivity and higher probability of error. Figure 3.7 suggests that both methods tend to produce similar results in detection performance as a function of vertex degree. Given this similarity, one should opt for the reduced complexity of using single penalty with tuning parameter selected by (3.30).

3.6 Conclusion

We have presented an estimator of the topology of interactions in a spatio-temporal graphical model. While the penalized likelihood formulation was derived



a.) b.) c.)

Figure 3.6: % zeros in the reconstruction of edges 200 node synthetic small world network under 1000 time points resampled over 1000 initial conditions of 40 randomly selected nodes as infected with rest susceptible. a.) ground truth, b.) single tuning parameter, c.) multiple tuning parameters (white - 0% black 100%)

Table 3.4: Detection statistics vs. time horizon for 200 node synthetic small-world network with trajectories resampled over 1000 initial conditions of 40 randomly selected nodes as "infected" with rest "susceptible"

Method	T	$Sens.(\lambda^*)$	$Spec.(\lambda^*)$	$P_e(\lambda^*)$
ℓ_1-SIR(λ^*)	100	0.26(0.05)	0.94(0.01)	0.08(0.01)
ℓ_1-SIR($\{\lambda_i^*\}_{i=1}^p$)	100	0.28(0.04)	0.92(0.01)	0.10(0.01)
ℓ_1-SIR(λ^*)	400	0.41(0.02)	0.95(0.00)	0.07(0.00)
ℓ_1-SIR($\{\lambda_i^*\}_{i=1}^p$)	400	0.46(0.02)	0.93(0.00)	0.08(0.00)
ℓ_1-SIR(λ^*)	700	0.77(0.02)	0.90(0.01)	0.11(0.01)
ℓ_1-SIR($\{\lambda_i^*\}_{i=1}^p$)	700	0.77(0.02)	0.90(0.00)	0.11(0.00)
ℓ_1-SIR(λ^*)	1000	0.87(0.01)	0.90(0.00)	0.07(0.01)
ℓ_1-SIR($\{\lambda_i^*\}_{i=1}^p$)	1000	0.87(0.02)	0.90(0.00)	0.07(0.00)

for the general SIR model, more complex SIR processes, *i.e.*, $SI_1 \cdots, I_m RS$ could be handled by our approach. The detection performance resulting from simulations of a H1N1 epidemic model suggests that the proposed method accurately reconstructs the topology of these types of networks while outperforming other state of the art structure learning algorithms.

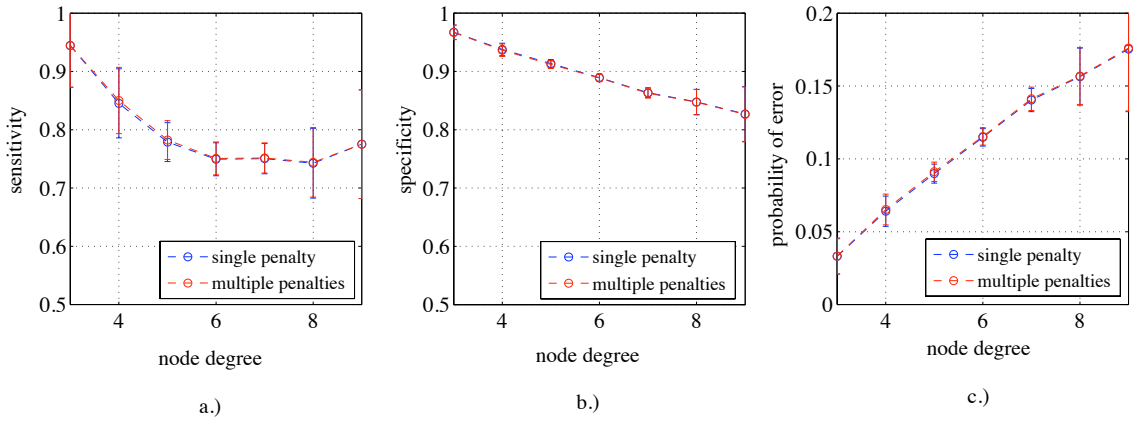


Figure 3.7: Neighborhood detection statistics vs. node degree for 200 node small-world network with $T = 1000$ with trajectories resampled over 1000 initial conditions of 40 randomly selected nodes as infected with rest healthy. a.) sensitivity, b.) specificity, c.) probability of error (red Single Penalty, blue Multiple Penalties)

CHAPTER IV

Robust Logistic Regression with Bounded Data Uncertainties

4.1 Introduction

There are two common methods for accommodating uncertainty in the observed data in risk minimization problems. The first approach assumes stochastic measurement corruption, centered about the true signal. This method is commonly known as error in variables (EIV) and has a rich history in least-squares and logistic regression (LR) problems [62, 60, 5, 26]. Unfortunately, EIV estimators are optimistic, require solving non-convex optimization problems, and de-regularize Hessian-like matrices making numerical estimation less stable. The latter approach of accommodating measurement uncertainty involves developing estimators that are robust to worst-case perturbations in the data and result in solving well posed convex programs [7, 20, 15, 31, 28, 63, 2].

The work presented throughout this paper builds on the results of [15] and [7]. Here, we generalize the robust optimization problem to a variety of different uncertainty sets appropriate for real problems. We present thresholding conditions which produce block-sparse parameters when confronted with grouped uncertainty. The robust risk functions, resulting from the minimax estimation, are regularized with group structured penalties to accommodate high dimensional data when the

underlying signal is both block-sparse and measurements are uncertain. A block co-ordinate gradient descent with an active shooting speed up algorithm is presented which exploits the iterative grouped thresholding conditions. The relationship between ridge LR and robust LR (RLR) is discussed. The robustness of ridge LR is established by identifying conditions when the uncertainty magnitude of robust can be re-parameterized in terms of the ridge tuning parameter such that both methods yield the same solution. Conditions on the Hessians of each method are established such that the RLR approach converges to this solution faster than ridge LR. We also present an empirical approach to estimating the uncertainty bounds using quantiles. We conclude by presenting a motivating example using gene expression data and discuss how robust ℓ_1-regularization paths recover "robust genes" that were previously over looked by standard ℓ_1-regularization paths. The worst-case probability of errors and false alarm rates of RLR are always less than or equal to those from LR. The results suggest that "robustification" of logistic classifiers can lead to significant performance gains in gene expression analysis.

4.2 Robust Logistic Regression

The goal of robust logistic regression (RLR) is to extract an estimator by minimizing the worst-case errors in measurements on the LR loss function (binomial deviance) subject to bounded uncertainty. The general case of RLR with spherical uncertainty, previously explored by [15], involves n measured training variables $\{x_i\}_{i=1}^n$, $x_i \in \mathbb{R}^p$ with minimax formulation

$$\min_{\beta,\beta_0} \max_{\{\delta_i\}_{i=1}^n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i\left(\beta^T(x_i+\delta_i)+\beta_0\right)}\right)$$

$$\text{subject to} \quad \|\delta_i\|_{\ell_2} \leq \rho \;\forall\; i = 1, \ldots, n. \tag{4.1}$$

Here, the true signals are given by $\{z_i = x_i + \delta_i\}_{i=1}^n$ and the parameter ρ is the magnitude of the worst-case perturbation. The minimax problem in (4.1) is solved by first solving the inner maximization step analytically, resulting in a convex RLR loss function which is then minimized with respect to β, β_0. We will confine our efforts to the situation in which class labels are perfectly observed.

Note that the maximization over each of the n perturbations, δ_i can be moved within the sum loss function in (4.1)

$$\min_{\beta,\beta_0} \sum_{i=1}^{n} \max_{\delta_i} \log\left(1 + e^{-y_i\left(\beta^T(x_i+\delta_i)+\beta_0\right)}\right)$$

$$\text{subject to } \|\delta_i\|_{\ell_2} \leq \rho \;\forall\; i = 1, \ldots, n. \tag{4.2}$$

We would like to reduce the minimax problem in (4.2) to a closed form minimization problem over β as performed in [7]. We begin by noting the following

$$-y_i\beta^T\delta_i \leq \|\beta\|_{\ell_2}\|\delta_i\|_{\ell_2} \leq \|\beta\|_{\ell_2}\rho. \tag{4.3}$$

Given that the loss function is monotonic in $-y_i\beta^T\delta_i$, we have the following upper-bound on the binomial deviance for the i^{th} sample:

$$\log\left(1 + e^{-y_i\left(\beta^T(x_i+\delta_i)+\beta_0\right)}\right) \leq \log\left(1 + e^{-y_i\left(\beta^T x_i+\beta_0\right)+\rho\|\beta\|_{\ell_2}}\right). \tag{4.4}$$

The upper-bound in both (4.3) and (4.4) is achievable for δ_i collinear with β, i.e., $\delta_i = \gamma_i\beta$ for some alignment parameter γ_i, thus yielding the solution to the constrained maximization step for the i^{th} observation

$$\max_{\delta_i, \|\delta_i\|_{\ell_2}\leq\rho} \log\left(1 + e^{-y_i\left(\beta^T(x_i+\delta_i)+\beta_0\right)}\right) = \log\left(1 + e^{-y_i\left(\beta^T x_i+\beta_0\right)+\rho\|\beta\|_{\ell_2}}\right). \tag{4.5}$$

We may now proceed with obtaining the robust estimated normal vector β corresponding to the binomial deviance via the following unconstrained minimization

problem:

$$\min_{\beta,\beta_0} \sum_{i=1}^{n} \log\left(1 + e^{-y_i\left(\beta^T x_i + \beta_0\right) + \rho\|\beta\|_{\ell_2}}\right). \tag{4.6}$$

Geometrically, the robust binomial deviance penalizes points based upon their distance to the "uncertainty margins" $\pm\rho\|\beta\|_2$, i.e., a point is penalized more for being the same distance away from the hyperplane under the robust formulation than standard LR. This pessimism is intuitive as observations that are close to the decision boundary could have their true value lying on the misclassified side under worst-case perturbations (see Figure 4.1).

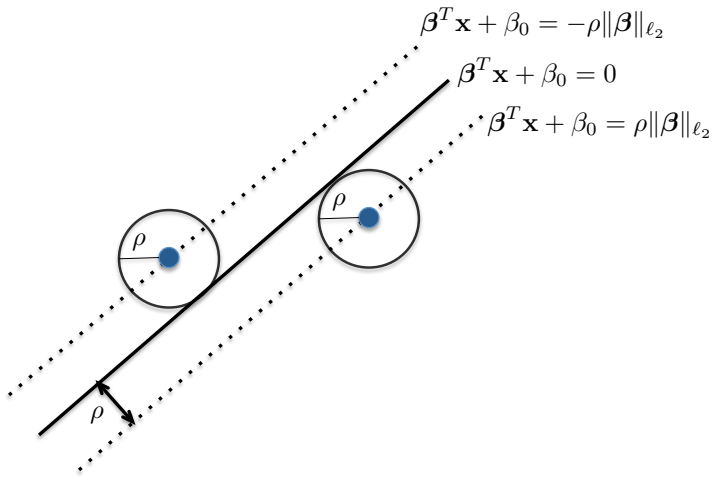


Figure 4.1: Bounded Uncertainty Modification Penalizes Based on *Potentially* Mis-Classified Points Translating Logistic Regression Loss to Penalize Based on Margins

4.2.1 Robust Logistic Regression with Group Structured Uncertainty Sets

In practice, joint spherical uncertainty may be inappropriate for modeling real data. A more appropriate form of uncertainty occurs when it affects groups of variables. Here, we assume that perturbations have group structure and are applied to G disjoint subsets of the p variables. These assumptions produce the following

robust optimization problem

$$\min_{\beta,\beta_0} \sum_{i=1}^{n} \max_{\delta_i} \log\left(1 + e^{-y_i\left(\beta^T(x_i+\delta_i)+\beta_0\right)}\right)$$

$$\text{subject to } \{\|\delta_{i,g}\|_{\ell_2} \leq \rho_g\}_{g=1}^{G} \ \forall\ i = 1, \ldots, n \tag{4.7}$$

where $\delta_i = \{\delta_{i,g}\}_{g=1}^{G}$ with $\delta_{i,g} \in \mathbb{R}^{|\mathcal{I}_g|}$ and $\mathcal{I}_g \subset \mathcal{I}$, $\mathcal{I} = \{1, \ldots, p\}$, are the set indices corresponding to the variables in group g. We will assume that $\mathcal{I}_g \cap \mathcal{I}_{g'} = \emptyset$ for $g \neq g'$. Note that that loss function is monotonic in $\sum_{g=1}^{G} -y_i\beta_g^T\delta_{i,g}$, and therefore, under worst-case perturbations, we have

$$\sum_{g=1}^{G} -y_i\beta_g^T\delta_{i,g} \leq \sum_{g=1}^{G} \|\beta_g\|_{\ell_2}\|\delta_{i,g}\|_{\ell_2} \leq \sum_{g=1}^{G} \rho_g\|\beta_g\|_{\ell_2}. \tag{4.8}$$

The inner maximization step in (4.7) is achieved when the upper bounds in (4.8) is tight, which is when β_g is colinear with $\delta_{i,g}$. Therefore, as in (4.5), we have analytically computed the inner-maximization step, and thus, our problem reduces to solving the following:

$$\min_{\beta,\beta_0} \sum_{i=1}^{n} \log\left(1 + e^{-y_i\left(\beta^T x_i+\beta_0\right)+\sum_{g=1}^{G}\rho_g\|\beta_g\|_{\ell_2}}\right). \tag{4.9}$$

The term within the argument of the RLR loss function is the "group lasso" penalty (4.9) which tends to promotes spareness in the groups (or factors) when used to regularize convex risk functions [66, 36].

When the number of groups is $G = p$ (each variable in its own group), the perturbations are interval based, and the group lasso penalty is equivalent to the ℓ_1-penalty, and thus, our minimization problem becomes when $\rho_g = \rho$, for all $g = 1, \ldots, p$

$$\min_{\beta,\beta_0} \sum_{i=1}^{n} \log\left(1 + e^{-y_i\left(\beta^T x_i+\beta_0\right)+\rho\|\beta\|_{\ell_1}}\right). \tag{4.10}$$

The minimization problem in (4.10) was previously treated in the context of interval perturbations in [15].

4.2.2 Regularized Robust Logistic Regression

Penalties such as the ℓ_1-norm are used in high-dimensional data settings, as they tend to zero out many of elements in β, which may better represent the structure of the underlying signal. Many fields of research increasingly involve high-dimensional data measurements that are obtained under noisy measurement conditions, such as gene expression microarrays that measure the activity of thousands of genes by assaying the abundances of mRNA in the sample. Here we develop new logistic classifiers that have the combined advantages of sparsity in variables and robustness to measurement uncertainty.

Here, we will assume that the group structure of the regularization penalties coincide with the structure of uncertainty sets. For an arbitrary set of G disjoint groups, the following regularized robust solution is

$$\min_{\beta_g,\beta_0} \sum_{i=1}^{n} \log\left(1 + e^{-y_i(\beta^T x_i+\beta_0)+\sum_{g=1}^{G}\rho_g\|\beta_g\|_{\ell_2}}\right) + \sum_{g=1}^{G}\lambda_g\|\beta_g\|_{\ell_2}. \tag{4.11}$$

The presence of the additional group-lasso penalty should promote block-sparsity in β while being robust to measurement error affecting the same variables in the group.

4.3 Computations for Regularized Robust Logistic Regression

Here, we present a numerical solution to the general regularized RLR problem based on block co-ordinate gradient descent with an active shooting step to speed up the convergence time when confronted with sparse signals or many groups. Since the penalized loss function in (4.11) is convex, denoted by $L_{\rho,\lambda}$, we can iteratively obtain β_g via block coordinate gradient descent. However, the gradient of (4.11) does not exist at $\beta_g = 0$, and thus we must resort to sub-gradient methods to identify optimality conditions ([3, 36, 52]). The necessary and sufficient conditions for β_g to

be a valid solution of (4.11) require

$$-X_g^T A_\rho y + (\rho_g \text{ tr}\,(A_\rho) + \lambda_g)\,\frac{\beta_g}{\|\beta_g\|_{\ell_2}} = 0\,, \beta_g \neq 0$$
$$\|X_g^T A_{-\beta_g,\rho} y\|_{\ell_2} \leq \left(\rho_g \text{tr}(A_{-\beta_g,\rho}) + \lambda_g\right), \beta_g = 0$$

with $A_\rho = (I + K_\rho)^{-1}$, $K_\rho = \text{diag}\left(e^{y_i\left(\beta^T x_i + \beta_0\right) - \sum_{g=1}^{G} \rho_g \|\beta_g\|_{\ell_2}}\right)$. Note that $A_{-\beta_g,\rho}$ means that β_g is set to 0 in K_ρ. The group thresholding that arises from these optimality conditions appears in group lasso regularized problems [66, 36, 52] and to the authors knowledge, has not been previously extracted in the context of RLR [15]. While RLR uses a different loss function than the binomial deviance, the thresholding conditions (above) establish the relationship between regularizing a standard convex risk function with a non-differentiable penalty and the sparse solutions that tend to appear when using robust linear classifiers with uncertain data [15]. It is intuitive that the thresholding conditions depend on both the uncertainty magnitude ρ_g and the sparseness penalty parameter λ_g.

The proposed block co-ordinate gradient descent consists of updating the g^{th} group parameters by initially computing a Newton-step

$$\delta\beta_g^{(m)} = -\left[\nabla^2_{\beta_g} L^{(m)}_{\rho,\lambda}\right]^{-1} \nabla_{\beta_g} L^{(m)}_{\rho,\lambda} \tag{4.12}$$

followed by performing a backtracking line search ([4]) for appropriate step size $\nu^{(m)} > 0$, and then updating $\beta_g^{(m+1)}$

$$\beta_g^{(m+1)} \leftarrow \beta_g^{(m)} + \nu_g^{(m)} \delta\beta_g^{(m)}. \tag{4.13}$$

The numerical solution to solving (4.11) is outlined below in Algorithm 1. The active shooting [46] steps updates the parameters that were non-zero after the initial step until convergence. After this subset has reached convergence, then gradient descent is performed over all the variables. This preferential update tends to reach the global minima faster when confronted with many groups or sparse signals.

Algorithm 2: Active Shooting Block Coordinate Descent with Group Thresholding

1. Initialize:
   (a) $\beta_0^{(1)} \leftarrow \nu_0^{(0)} \delta\beta_0^{(0)}$ with all parameters set to zero
   (b) $\beta_g^{(1)} \leftarrow \nu_g^{(0)} \delta\beta_g^{(0)}$, for $g = 1, \ldots, G$, with β_0 evaluated at $\beta_0^{(0)}$ and all other parameters set to zero
2. Define the active set $\Lambda = \{g : \beta_g^{(0)} \neq 0\}$
3. $\beta_0^{(m+1)} \leftarrow \beta_0^{(m)} + \nu_0^{(m)} \delta\beta_0^{(m)}$ with $\delta\beta_0^{(m)}$ via (4.12), $\nu_0^{(m)}$ by performing backtracking, and β held at previous value
4. For $g \in \Lambda$
   (a) if $\|X_g^T A_{-\beta_g,\rho} y\|_{\ell_2} \leq \rho_g \text{tr}\left(A_{-\beta_g,\rho}\right) + \lambda_g$, $\beta_g^{(m+1)} \leftarrow 0$
   (b) else, evaluate $\delta\beta_g^{(m)}$ from (4.12) while holding all other parameters at previous values, compute step size $\nu_g^{(m)}$ via backtracking, and update $\beta_g^{(m+1)} \leftarrow \beta_g^{(m)} + \nu_g^{(m)} \delta\beta_g^{(m)}$
5. Repeat steps 3 and 4 until some convergence criteria met for active parameters in Λ.
6. If convergence criteria satisfied, define $\Lambda = \{1, \ldots, G\}$ and repeat 3 and 4 until convergence in all parameters.

4.4 Empirical Estimation of Uncertainty

The magnitude of the potential uncertainty is determined by ρ_g. There are situations when a researcher has prior knowledge on the value of ρ_g but more often this parameter must be estimated empirically. In modern biomedical experiments,

in which gene expression microarrays are used to assay the activity of tens of thousands of genes, technical replicates are frequently obtained to assess the effect of of measurement uncertainty.

We will estimate ρ_g using a generalization of the method in [61] based on quantiles. For grouped uncertainty sets, we estimate ρ_g via the following

$$\hat{\rho}_g(\alpha) = \inf_\tau \mathbb{P}(\sqrt{x_g^T x_g} \leq \tau) = \alpha \tag{4.14}$$

where the distribution in (4.14) is taken with respect to a data set independent of the training data, such as technical replicates of a biological experiment. Note that (4.14) reduces to interval based quantile estimates when the number of groups $G = p$. As the cumulative distribution function (CDF) in (4.14) does not depend on class label y, we obtain (4.14) by

$$\mathbb{P}(z_g \leq \tau) = \sum_{y \in \{-1,+1\}} \mathbb{P}(z_g \leq \tau | Y = y)\mathbb{P}(Y = y) \tag{4.15}$$

with $z_g = \sqrt{x_g^T x_g}$ and data centered about their respective class centroids. The class priors are estimated empirically by $\hat{\mathbb{P}}(Y = y) = m_y/m$, where m_y and m are the number of replicate samples with label y and total number of replicate samples, respectively. The estimation in (4.14) can be assessed with respect to the empirical CDF of the data $\{x_{i,g}\}_{i=1}^n$ or approximated by the inverse-CDF of the χ_{p_g} distribution [61] with $p_g = |\mathcal{I}_g|$ degrees of freedom. The application of the proposed estimation of uncertainty bounds is presented within the results section in the context of high dimensional gene expression data in which technical replicates are available.

4.5 Robust vs. Ridge Regression

4.5.1 Robustness of Ridge Logistic Regression

One important question is how the proposed RLR formulation relates to Ridge LR. Ridge LR involves adding a squared ℓ_2-penalty to the binomial deviance loss

function:

$$\min_\beta \sum_{i=1}^{n} \log\left(1 + e^{-y_i \beta^T x_i}\right) + \frac{\lambda}{2}\|\beta\|_{\ell_2}^2. \tag{4.16}$$

Our goal is to identify values of ρ as a function of λ for which both robust (4.1) and ridge (4.16) produce approximately identical estimates of β. We begin by inspecting the optimality conditions for β. The gradients for ridge and robust, respectively, are given as (intercept removed for clarity):

$$\nabla_\beta L_\lambda = -X^T W y + \lambda\beta \tag{4.17}$$

$$\nabla_\beta L_\rho = -X^T W_\rho y + \rho \mathrm{tr}(W_\rho)\frac{\beta}{\|\beta\|_{\ell_2}}. \tag{4.18}$$

where $W = \mathrm{diag}(\frac{e^{-y_i\beta^T x_i}}{1+e^{-y_i\beta^T x_i}})$ and $W_\rho = \mathrm{diag}(\frac{e^{-y_i\beta^T x_i + \rho\|\beta\|_{\ell_2}}}{1+e^{-y_i\beta^T x_i+\rho\|\beta\|_{\ell_2}}})$. If we linearize the sigmoidal terms in W and W_ρ, the scaled gradients can be approximated by the following:

$$\begin{aligned} n^{-1}\nabla_\beta L_\lambda &\approx \left(\frac{1}{4n}X^T X + (\lambda/n)I\right)\beta - \frac{1}{2}\delta\bar{x}_n \\ &= H_{n,\lambda}\beta - \frac{1}{2}\delta\bar{x}_n \end{aligned} \tag{4.19}$$

with $\delta\bar{x}_n = n^{-1}\left(n_+\bar{x}_+ - n_-\bar{x}_-\right)$

$$\begin{aligned} n^{-1}\nabla_\beta L_\rho &\approx \frac{1}{4n}X^T X\beta - \frac{1}{2}\delta\bar{x}_n + \frac{1}{4}\rho^2\beta \\ &+ \frac{\rho}{2\|\beta\|_{\ell_2}}\left((1 - \frac{\beta^T\delta\bar{x}_n}{2})\beta - \frac{1}{2}\|\beta\|_{\ell_2}^2\delta\bar{x}_n\right) \\ &= c_\beta + a_\beta\rho^2 + b_\beta\rho. \end{aligned} \tag{4.20}$$

Since the gradient for ridge regression can be approximated by the linear system of equations in (4.19), we can approximate β_λ via

$$\hat{\beta}_\lambda \approx \frac{1}{2}H_{n,\lambda}^{-1}\delta\bar{x}_n. \tag{4.21}$$

Inserting (4.21) into (4.20), summing (4.20) over its elements by an inner product with a vector of 1's, and solving for ρ, produces the following quadratic equation of ρ in terms of λ

$$\rho_\lambda \approx \frac{-1^T b_{\hat{\beta}_\lambda} \pm \sqrt{\left(1^T b_{\hat{\beta}_\lambda}\right)^2 - 4 \cdot 1^T a_{\hat{\beta}_\lambda} \cdot 1^T c_{\hat{\beta}_\lambda}}}{2 \cdot 1^T a_{\hat{\beta}_\lambda}}. \tag{4.22}$$

The positive value of ρ_λ is the uncertainty magnitude. This establishes equivalence between the solutions of ridge and robust logistic regression for a given λ.

4.5.2 Convergence Rates of Ridge and Robust

We established conditions of equivalence between ridge and robust LR estimators through a re-parameterization of ρ in terms of λ (4.22). It is also of interest to investigate the rate at which these methods converge under the proposed conditions of equivalence. For ease of exposition, we will assume that the data has been centered and there are balanced sample sizes ($n_+ = n_-$). We will assume that the data is not "too far" from the decision boundary (as stated above).

We begin by assuming that the gradients of both ridge LR (4.19) and RLR (4.20) are in some neighborhood about 0. The structure of the Hessian specify rates of convergence through the eigenvalues. Under the assumptions detailed above, the scaled Hessian for the ridge case is obtained from differentiating (4.19) is given by

$$n^{-1}\nabla^2_\beta L_\lambda \approx \frac{1}{4n} X^T X + (\lambda/n) I. \tag{4.23}$$

By differentiating (4.20) and noting that $\frac{1}{8}\beta^T \delta\bar{x}_n \ll \frac{1}{2}$, we obtain the scaled Hessian for robust logistic regression

$$n^{-1}\nabla^2_\beta L_\rho \approx \left(\frac{1}{4n} X^T X + \frac{1}{4}\rho^2 I\right) + \rho\,(A - B) \tag{4.24}$$

with positive semi-definite matrix A

$$A = \frac{1}{2\|\beta\|_{\ell_2}} \left(I - \frac{1}{\|\beta\|^2_{\ell_2}} \beta\beta^T\right) \tag{4.25}$$

and positive semi-definite matrix B

$$B = \frac{1}{8\|\beta\|_{\ell_2}} \delta\bar{x}_n \beta^T. \tag{4.26}$$

We are interested in extracting conditions on ρ such that the Hessian corresponding to the robust binomial deviance is "larger" than that corresponding to ridge, i.e., $\Delta H = \nabla^2_\beta L_\rho - \nabla^2_\beta L_\lambda \succeq 0$, when both methods yield the same β, or equivalently

$$w^T \Delta H w \geq 0, \forall w. \tag{4.27}$$

One can directly derive necessary conditions that must be satisfied by ΔH by enforcing that (4.27) holds for some choice of w. We take $w = \delta\bar{x}$, which is the right eigenvector of B, and substitute $\hat{\beta}_\lambda$ (4.21) and ρ_λ (4.22) for β and ρ, respectively in (4.27). The conditions such that robust logistic regression converges faster to the solution of ridge logistic regression are

$$\rho_\lambda^2 + \frac{2(1-\epsilon'_\lambda)}{\|\hat{\beta}_\lambda\|_{\ell_2}} \rho_\lambda \geq 4\lambda/n \tag{4.28}$$

where ϵ'_λ is given by

$$\epsilon'_\lambda = \left(\hat{\beta}_\lambda^T \delta\bar{x}\right)^2 \left(\frac{1}{\|\delta\bar{x}\|^2_{\ell_2} \|\hat{\beta}_\lambda\|^2_{\ell_2}} + \frac{1}{4}\right). \tag{4.29}$$

4.6 Numerical Results

4.6.1 Recovery of Regularization Path Under Signal Corruption

Many researchers in machine learning and bioinformatics assess importance of various biomarkers based upon the ordering of ℓ_1-regularization paths. Thus, it is worthwhile to explore how the ordering of the variables within the regularization path change as uncertainty is added. We first show a simple synthetic example there are $n = 100$ samples in each class with $x \in \mathbb{R}^p$, $p = 22$, with $x \sim \mathcal{N}(\mu_y, \Sigma)$. The class centroids are given by $\mu_{+1} = [-\frac{1}{2}, \frac{1}{3}, -\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, 0, \cdots, 0]^T$ and $\mu_{-1} =$

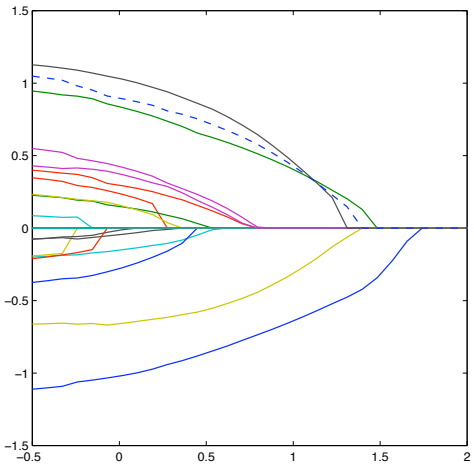
$[\frac{1}{2}, -\frac{1}{3}, 0, \cdots, 0, 0, 0, 0, 0]^T$. The structure of Σ is block diagonal, where diagonal components $\{\sigma_{ii}\}_{i=1}^{p}$ are equal to one and the block structure affects only the variables $x_1, x_2, x_3, x_4, x_5, x_6$, where the off-diagonal elements in this block $\{\sigma_{ij}\}_{j=1, j\neq i}^{6}$ are equal to 0.1. The other off-diagonal elements have covariance of zero.

Figure 4.2(a) represents the ℓ_1-regularization path as a function of $\log_{10}\lambda$ on the original data. In Figure 4.2(b), x_3 has been perturbed such that $x_3 \leftarrow x_3 + \delta_3$ with $-\rho \leq \delta_3 \leq \rho$, which leads to a shift in the ordering within the regularization path under normal ℓ_1-LR. Using the perturbed data set and knowledge that $\rho = 0.1$ for x_3 and presented with interval uncertainty, the ℓ_1-regularized robust logistic regression recovers the original ordering of the variables (see Figure 4.2(c)).

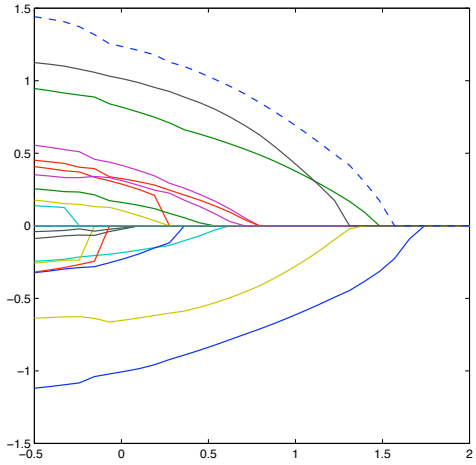
4.6.2 Human Rhino Virus Gene Expression Data

Here we present numerical results on peripheral blood gene expression data set sampled over 14 time points from a group of $n = 20$ patients inoculated with the Human Rhino Virus (HRV), the typical agent of the common cold [67]. Half of the patients responded with symptoms ($y = +1$) and the other half did not ($y = -1$). The original $12,023$ genes on the microarray were reduced to $p = 129$ differentially expressed genes controlling for a 20% False Discovery Rate [54]. We will regularize both the robust binomial deviance and standard binomial deviance with an ℓ_1-penalty to control the sparsity of the resulting model forcing many elements in β to be zero. In this experiment there were approximately 20 microarray chips that were technical replicates. The technical replicates were used to estimate the interval uncertainty bounds using (4.14).

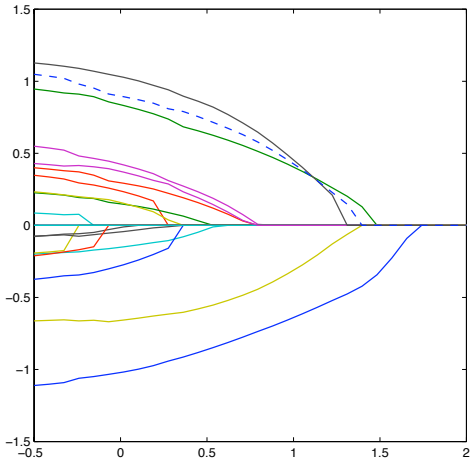
Since microarray devices detect the presence of mRNA abundance by including thousands of different probe sets (short sequences) that bind to a particular mRNA molecule, produced from a specific gene, we will adopt interval uncertainty across



(a) ℓ_1-Regularization Path on Original Data



(b) ℓ_1-Regularization Path with Bounded Perturbations in x_3 and $\rho_3 = 0.1$



(c) ℓ_1-Robust Regularization Path with $\rho_3 = 0.1$

Figure 4.2: Regularization Paths as a Function of $\log_{10}\lambda$: Robust Recovers Original Ordering After Perturbation

the p genes. The estimated interval uncertainty bounds, $\{\hat{\rho}_i(\alpha)\}_{i=1}^p$, were obtained from (4.14) using the empirical CDF of the independent technical replicate data with linear interpolation between sampled data. The CDF in (4.14) was computed by (4.15) with each conditional CDF centered about their class dependent sample mean and equal class priors $\hat{\mathbb{P}}(Y = +1) = \hat{\mathbb{P}}(Y = -1) = 1/2$. Technical replicates are generated from the blood sample used to assess gene expression activity in the training data and thus, are commonly used to isolate the effect of measurement error.

We explored the effect of interval uncertainty bounds resulting from quantiles of (4.14) at levels of $\alpha = \{0.25, 0.50, 0.75, 0.95, 0.99\}$. Figure 4.3 shows the ℓ_1-regularization paths obtained by solving (4.11) on the training data for different interval uncertainty (including no uncertainty corresponding to standard ℓ_1-LR) and illustrates how robustness affects the ordering of the genes. We see from Figure 4.3(a) that the first gene to appear in the regularization path is the anti-viral defense gene RSAD2. RSAD2 persists at the first gene in the regularization path for $\alpha = \{0, 0.25, 0.50, 0.75\}$ (latter three not shown) but disappears from the regularization path completely, along with a few other genes, in Figure 4.3(a), when $\alpha = \{0.95, 0.99\}$. The three "robust genes" that persist across all explored values of α are ADI1, OAS1, and TUBB2A. Of these three, OAS1 (codes for proteins involved in the innate immune response to viral infection) barely appears in Figure 4.3(a), yet is the first gene to appear in the robust regularization paths that is common to both $\alpha = \{0.95, 0.99\}$ (see Figures 4.3(e) and 4.3(f)). These results suggest that when assessing variable importance via regularization paths, one should be aware of the effect of measurement uncertainty on the ordering of the variables.

The robust formulation within this paper aims at minimizing the worst-case configuration of perturbations on the logistic regression loss function. As our goal is discriminating between two phenotypes, it is of interest to compare the worst-case probability of error for LR and RLR on perturbed data sets, i.e., $x_i \leftarrow x_i + \delta_i$, $|\delta_i| \leq \hat{\rho}_i(\alpha)$, after training on the original data. The ℓ_1-tuning parameter λ for both standard ℓ_1-LR and ℓ_1-RLR was chosen to minimize the out-of sample probability of error via 5-fold cross validation. Given the cross validated value of λ for both methods, the models were then fit to the entire set of training data. 50,000 perturbed data matrices were then generated subject to interval uncertainty bounds $\{\hat{\rho}_i(\alpha)\}_{i=1}^p$

for $\alpha = \{0.25, 0.50, 0.75, 0.95, 0.99\}$. The best-case and worst-case probability of error were recorded in Table 4.1. We see for all values of α explored, the worst-case probability of error corresponding to ℓ_1-RLR is always less than or equal to that of ℓ_1-LR. When $\alpha = \{0.95, 0.99\}$, the worst-case probability of error for ℓ_1-LR is 0.30 but reduces to 0.20 when using ℓ_1-RLR with interval uncertainty estimated from the data using (4.14). Corresponding to the worst-case probability of errors are the sensitivity and 1-specificity, given in Table 4.2. We see that ℓ_1-RLR achieves the same power as ℓ_1-LR at false alarm rates less than or equal to that of the non-robust method. The proposed method can be used to reduce classifier error sensitivity in large scale classification problems. This can be important in practical applications such as biomarker discovery and predictive health and disease.

Table 4.1: Best and Worst Case Probability of Error, P_e, for the HRV Data Set

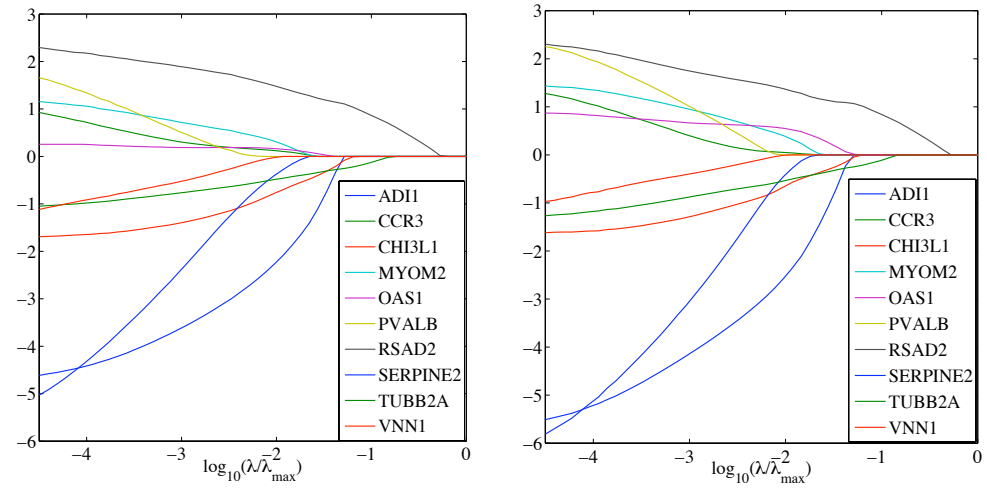
α	ℓ_1-LR		ℓ_1-Robust LR	
	min P_e	max P_e	min P_e	max P_e
0.25	0.10	0.15	0.10	0.15
0.50	0.05	0.20	0.10	0.20
0.75	0.10	0.25	0.10	0.25
0.95	0.00	0.30	0.05	0.20
0.99	0.00	0.30	0.00	0.20

Table 4.2: Sensitivity and 1-Specificity Corresponding to Worst Case Probability of Error from HRV Data

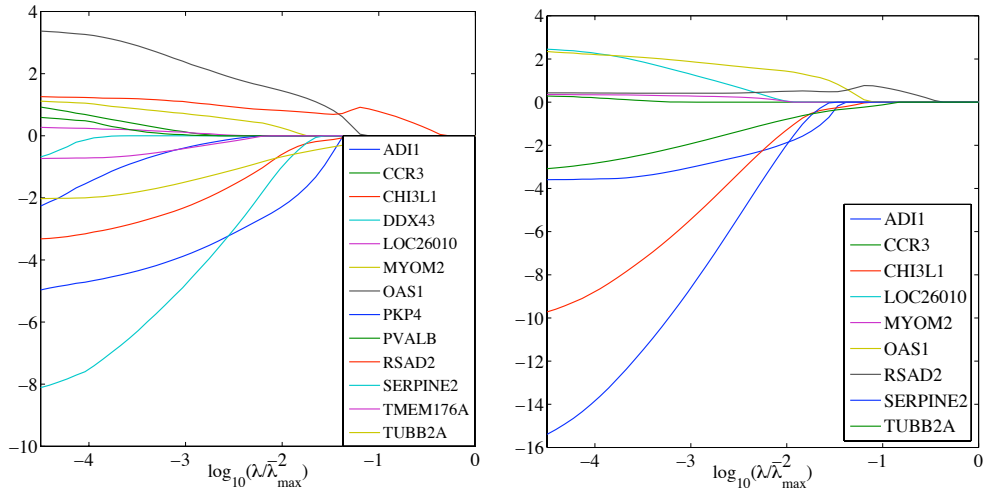
α	ℓ_1-LR		ℓ_1-Robust LR	
	Sens.	1-Spec.	Sens.	1-Spec.
0.25	0.70	0.00	0.70	0.00
0.50	0.70	0.10	0.70	0.10
0.75	0.70	0.20	0.70	0.20
0.95	0.70	0.20	0.70	0.10
0.99	0.70	0.30	0.70	0.10

4.7 Conclusion

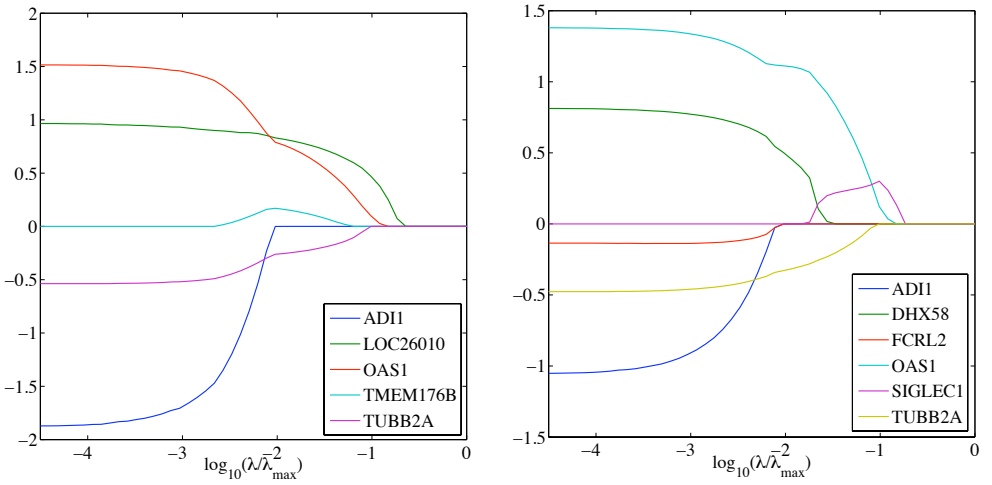
Building on the results of [15] and [7], we have formulated the robust logistic regression problem with group structured uncertainty sets. By adding regularization penalties, one can enforce block-sparsity assumptions of the underlying signal. We have presented a block co-ordinate gradient method with iterative grouped thresholding for solving the penalized RLR problems. The group thresholding is affected by both the group-lasso penalty and the magnitude of the worst-case uncertainty. Thus RLR tends to promote thresholding of highly uncertain variables, thus performing an initial step of variable selection. We have proposed an empirical approach to estimating the uncertainty using quantile estimation. This approach was applied to a real gene expression data set where quantile estimation was applied to a set of technical replicates. The numerical results on this real data set establish that our approach can yield lower worst-case probability of error and lower false alarm rates. Such a gain in worst-case detection performance can improve the performance for predictive health and disease, and in particular for predicting patient phenotype. It will be interesting to explore the situation when the group-structure of the uncertainty differs from that of the regularization penalty. Such a situation could potentially be solved by modifying the sparse group-lasso solution as in [17]. It is also of interest to explore the kernelization of this method in which one can study the propagation of bounded data uncertainty in a reproducing kernel Hilbert space.



(a) $\alpha = 0$ (b) $\alpha = 0.25$



(c) $\alpha = 0.50$ (d) $\alpha = 0.75$



(e) $\alpha = 0.95$ (f) $\alpha = 0.99$

Figure 4.3: Regularization paths on the HRV data set as a function of $\log_{10}\ \lambda/\lambda_{max}$ for different magnitudes of interval uncertainty, as determined by the α quantile

CHAPTER V

Functional Discriminants for Classification of High-Dimensional Time-Series

5.1 Introduction

This chapter treats the problem of performing classification given high-dimensional, discretely sampled, time-series data. We propose summarizing the discriminating power of each of the p variables by transforming the observed time-series to Gaussian Process (GP) [49, 34] log-odds ratios in similar vein as linear discriminant (LD) and quadratic discriminant (QD) analysis [23]. The transformation of the original p-dimensional time-series data to p different log-odds ratios represents the new basis used in forming a single powerful classifier using ℓ_1-Logistic Regression (ℓ_1-LR) [29, 45]. The proposed supervised learning framework naturally accommodates mis-alligned time-series (time-series with different time-stamps). The proposed framework is applicable to biomedical data analysis where sequential measurement sampling is used to discriminate between potential phenotypes or determine optimal treatment strategies. This work presented in this chapter develops a functional classifier given empirical high-dimensional time-series data.

The classification of multiple time-series is a difficult task. If a time-series is treated as a random vector and traditional classification techniques are applied, one is typically not capturing the causal nature of temporal data. To enforce such causal-

ity in the classification or regression model, researchers have proposed using the fused LASSO regularization penalty [56]. The fused LASSO penalizes the absolute deviance of the successive differences of the model parameters corresponding to each of the discretely sampled time-points. While this approach incorporates the direction of time into the resulting model, it requires aligned time-series, i.e., samples with identical time-stamps of identical length, a difficult constraint to enforce when gathering temporal biomedical data. GPs have been previously used as a functional regression tool for smoothing or prediction in time-series problems [49]. Recently, [34] has proposed summarizing the discretely sampled time-series as smooth mean functions in a Reproducing Kernel Hilbert Space (RKHS) where kernel based classification methods, such as Support Vector Machines (SVMs), can be used to discriminate between two classes which produces the data. Unfortunately, the method presented in [34] is designed for single feature problems rather than high-dimensional temporal data, such as temporal gene expression data. The proposed method represents the noisy time-series as functions and captures the discriminating behavior of such functions by summarizing then as log-odds ratios corresponding to GPs. These new basis functions are extracted for each of the p features/variables and then a single classifier is formed using ℓ_1-LR. The proposed method can be extended to accommodate the missing time-stamp dilemma by treating time as a random nuisance parameter and removing it using Bayesian methods. We exhibit the power of this method on a large human gene expression time-series data where the goal is discriminating between symptomatic and asymptomatic phenotypes given inoculation with Human Influenza A/H3N2 (H3N2), or Human rhino virus (HRV), or Human respiratory syncytial virus (RSV).

5.2 Gaussian Process Formulation

Gaussian Processes (GPs) are a functional extension to the multivariate Gaussian distribution [49]. A multivariate Gaussian distribution is fully specified by its mean and covariance matrix. When empirically estimating these quantities, one must estimate $T(T-1)/2$ parameters in the covariance matrix if T is the dimension of the variable. If T corresponds to the number of discretely sampled time points in a time-series, estimating these $T(T-1)/2$ may be statistically intractable. In addition to estimating the elements within the covariance matrix, the multivariate Gaussian density assumes a fixed dimension of the variable, T. When working with GPs, the covariance matrix is replaced with kernel functions which parameterize the covariance between multiple time-points. GPs are also not limited to fixed length time-series or alignment of the corresponding time-stamps. In other words, the likelihood of an out of sample test trajectory, of different length and different time-stamps can be naturally computed using a GP distribution with parameters estimated on empirical time-series of different dimension.

Here, we are confronted with n training trajectories of p features

$$\mathcal{D} = \{t_j, f_i^{(j)}, y_j\}_{j=1,i=1}^{n,p} \tag{5.1}$$

where $t_j = [t_{j_1}, \ldots, t_{j_k}]^T$ are the time stamps of the discretely sampled time-series in the j^{th} sample, $f_i^{(j)} = [f_i^{(j)}(t_{j_1}), \ldots, f_i^{(j)}(t_{j_k})]^T$ is the j^{th} measured trajectory corresponding to the i^{th} feature , and $y_j \in \{-1, +1\}$ is the class label. We will assume that the dynamics of the trajectories are conditionally dependent on label (phenotype). We will assume that these random functions can be decomposed into two independent stochastic terms, a GP and white measurement error:

$$f_i(t|y) = g_i(t|y) + \epsilon(t) \tag{5.2}$$

with signal represented by

$$g_i(t|y) \sim \mathcal{N}\left(\mu_{i,y}(t), k_{i,y}(t,t')\right) \tag{5.3}$$

and white measurements

$$\epsilon(t) \sim \mathcal{N}\left(0, \sigma_n^2 \delta_{t,t'}\right) \tag{5.4}$$

where $\delta_{t,t'}$ is the Dirac Delta function and σ_n^2 is the noise variance of the observation. Given the independence of (5.3) and (5.4), we can model our observed trajectories via

$$f_i(t|y) \sim \mathcal{N}\left(m_{i,y}(t), k_{i,y}(t,t') + \sigma_n^2 \delta_{t,t'}\right). \tag{5.5}$$

The index term y denotes the class dependent dynamics through the mean function $m_{i,y}(t)$ and covariance kernel function $k_{i,y}(t,t')$ relative to some reference time t'.

We are interested in exploiting the observed training trajectories to form the predictive posterior distribution of some new test observation evaluated at a time point t_k. Given the prior distributions defined in (5.2), the class dependent posterior distribution of $f_i(t_k|y)$ is given via

$$f_i(t_k|y)|\{t_j, f_i^{(j)}\}_{j:y_j=y} \sim \mathcal{N}\left(m_{i,y}(t_k), \Sigma_{i,y}(t_k)\right) \tag{5.6}$$

with posterior mean function

$$m_{i,y}(t_k) = k_{i,y}^T(t_k, t_y)\left(K_{i,y}(t_y,t_y) + \sigma_n^2 I\right)^{-1} f_{i,y} \tag{5.7}$$

and posterior covariance kernel

$$\Sigma_{i,y}(t_k) = \left(k_{i,y}(t_k,t_k) + \sigma_n^2\right) - k_{i,y}^T(t_k,t_y)\left(K_{i,y}(t_y,t_y) + \sigma_n^2 I\right)^{-1} k_{i,y}(t_k,t_y). \tag{5.8}$$

Here, $t_y = \{t_j\}_{j:y_j=y}$, $k_{i,y}(t_k,t_y)$ is the column vector of kernel evaluations between t_k and all time points in t_y, $K_{i,y}(t_y,t_y)$ is the matrix of kernel evaluations between all

time points in t_y, and $f_{i,y} = \{f_i^{(j)}(t_j)\}_{j:y_j=y}$ is the column vector of all measurements of the i^{th} feature corresponding to label y.

The distribution in (5.6) is the posterior distributions of $f_i(t_k|y)$ given a single test point t_k. The extension of (5.6) to a collection test time-points $t = [t_1, \ldots, t_m]^T$ is given by the following two distributions [49]:

$$f_i|t, \{t_j, f_i^{(j)}\}_{j:y_j=y} \sim \mathcal{N}\left(m_{i,y}(t), \Sigma_{i,y}(t)\right) \tag{5.9}$$

with posterior mean vector function

$$m_{i,y}(t) = K_{i,y}(t, t_y)\left(K_{i,y}(t_y, t_y) + \sigma_n^2 I\right)^{-1} f_{i,y} \tag{5.10}$$

and posterior kernel function matrix

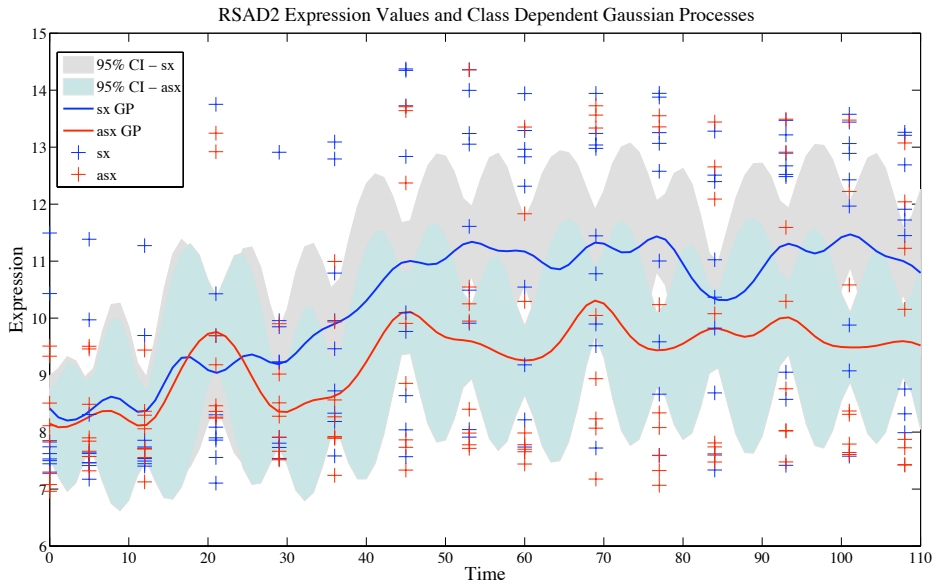
$$\Sigma_{i,y}(t) = \left(K_{i,y}(t,t) + \sigma_{i,n}^2 I\right) - K_{i,y}(t, t_y)\left(K_{i,y}(t_y, t_y) + \sigma_n^2 I\right)^{-1} K_{i,y}(t_y, t). \tag{5.11}$$

Figure 5.1 depicts the interpolation of the posterior mean function through observations enveloped with a 95% confidence interval of two differentially expressed genes, RSAD2 and IFI44, in a pan-viral human gene expression data set. We see at early hours (less than 30 hrs), that the symptomatic and asymptomatic mean functions are on top of eachother. From 30 hrs post-innoculation, we begin to see a divergence in the two mean functions, thus capturing the average divergence of the symptomatic class from the asymptomatic class.

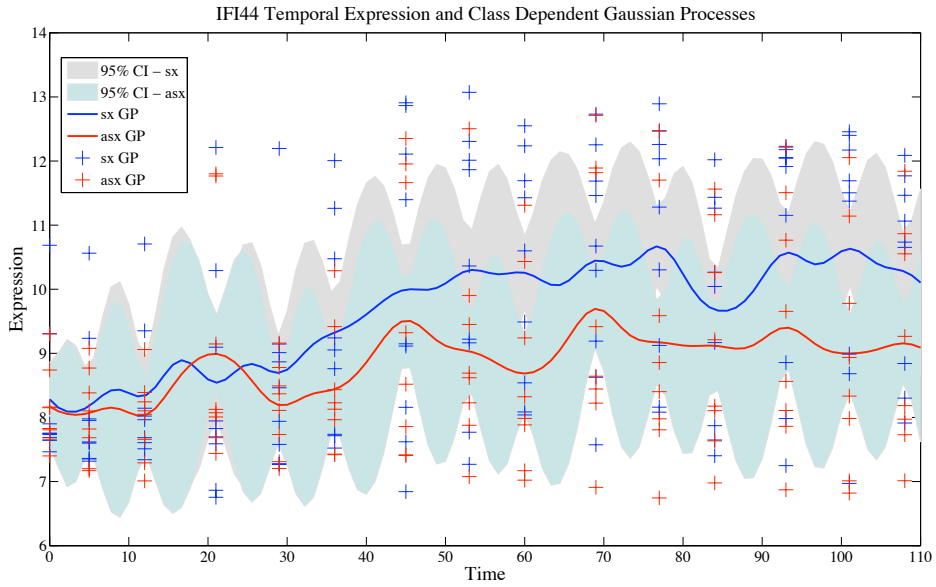
5.2.1 Functional Discriminants

Since we are interested in discriminating between different groups, Bayesian decision theory states that the posterior distribution of the class label is sufficient for optimal decision making [22]. Under our GP assumptions, we can form the posterior distribution of the class label given data for the i^{th} feature

$$\mathbb{P}\left(Y = y|\{t, f_i\}, \mathcal{D}_i\right) = \frac{\mathcal{N}\left(m_{i,y}(t), \Sigma_{i,y}(t)\right)\pi_y}{\sum_{y\in\{-1,+1\}} \mathcal{N}\left(m_{i,y}(t), \Sigma_{i,y}(t)\right)\pi_y} \tag{5.12}$$



(a) RSAD2



(b) IFI44

Figure 5.1: Phenotype Dependent Posterior GP Mean Function and 95% Confidence Intervals of Genes RSAD2 and IFI44 Resulting from Pan-Viral Human Challenge Study

with prior distribution $\pi_y = \mathbb{P}(Y = y)$ and training data for the i^{th} feature $\mathcal{D}_i = \{t_j, f_i^{(j)}, y_j\}_{j=1}^n$. When performing binary classification given a test trajectory $\{t, f_i\}$,

it is sufficient to inspect the odd-ratio, or equivalently the log-odds:

$$\begin{aligned}\phi_i(t, f_i) = \log \frac{\mathbb{P}\left(Y = +1 | \{t, f_i\}, \mathcal{D}_i\right)}{\mathbb{P}\left(Y = -1 | \{t, f_i\}, \mathcal{D}_i\right)} &= \log \frac{\mathcal{N}\left(m_{i,+1}[t], \Sigma_{i,+1}[t]\right)}{\mathcal{N}\left(m_{i,-1}[t], \Sigma_{i,-1}[t]\right)} + \log \frac{\pi_{+1}}{\pi_{-1}} \\ &= \frac{1}{2} f_i^T A_i(t) f_i + b_i^T(t) f_i + c_i(t) \end{aligned} \tag{5.13}$$

with

$$A_i(t) = \left(\Sigma_{i,-1}^{-1}(t) - \Sigma_{i,+1}^{-1}(t)\right) \tag{5.14}$$

$$b_i(t) = \left(\Sigma_{i,+1}^{-1}(t) m_{i,+1}(t) - \Sigma_{i,-1}^{-1}(t) m_{i,-1}(t)\right) \tag{5.15}$$

(5.16)

$$c_i(t) = \log \frac{\pi_{+1}}{\pi_{-1}} + \log \frac{|\Sigma_{i,-1}(t)|}{|\Sigma_{i,+1}(t)|} + m_{i,-1}^T(t) \Sigma_{i,-1}^{-1}(t) m_{i,-1}(t) - m_{i,+1}^T(t) \Sigma_{i,+1}^{-1}(t) m_{i,+1}(t).$$

We see from (5.13), that the functional quadratic classifier $\phi_i(t, f_i)$ is dependent on observed values of f_i and the corresponding observed time-points t, thus allowing a temporally dependent decision boundary, e.g., $\phi_i(t, f_i) = 0$, an important property of any classifier for temporal data.

We can collect each of the p basis functions (5.13) to form the p-dimensional column vector $\phi(t, f) = [\phi_1(t, f_1), \ldots, \phi_p(t, f_p)]^T$. We seek a single classifier that is an aggregate of all p log-odds such that the resulting classifier yields a lower probability of error than any single log-odds (5.13). Specifically, we seek the following classifier

$$h(t, f) = \beta^T \phi(t, f) + \beta_0 \tag{5.17}$$

with $\beta \in \mathbb{R}^p$ and bias controlling parameter $\beta_0 \in \mathbb{R}$. The parameters β, β_0 may be estimated by minimizing any convex risk function with possible added regularization penalties

$$\min_{\beta, \beta_0} \; L(\beta, \beta_0; \mathcal{D}) + \lambda J(\beta) \tag{5.18}$$

where $L(\beta, \beta_0; \mathcal{D})$ is a convex loss function, e.g., the logistic regression loss function, $\lambda > 0$ is a tuning parameter, and $J(\beta)$ is a convex penalty on β that enforces prior

knowledge, such as the ℓ_1-norm, which promotes a sparse $\hat{\beta}(\lambda)$ for sufficiently large λ, i.e., many $\hat{\beta}(\lambda)_i$'s are zero [55, 29, 36, 52].

The classifier $h(t, f)$ in (5.17) consists of a linear combination of the p log-odds, resulting from summarizing each of the p trajectories as log-odds ratios of GPs. Since many variables in stochastic systems jointly evolve over time it may be useful to add additional basis functions that capture second order effects. Specifically, we can model all $\binom{p}{2}$ pair-wise evolutions using the bivariate GP between features i and j

$$f_{ij}(t|y) = \begin{bmatrix} f_i(t|y) \\ f_j(t|y) \end{bmatrix} + \begin{bmatrix} g_i(t|y) \\ g_j(t|y) \end{bmatrix} + \begin{bmatrix} \epsilon(t) \\ \epsilon(t) \end{bmatrix} \tag{5.19}$$

with covariance kernel matrix

$$K_{ij,y}(t, t') = \begin{bmatrix} k_{i,y}(t, t') & k_{ij,y}(t, t') \\ k_{ij,y}^T(t, t') & k_{j,y}(t, t') \end{bmatrix}. \tag{5.20}$$

Under the bivariate GP modeling, the predictive posterior distribution of $f_{ij}(t|y)$ evaluated at a sequence of time points t is

$$f_{ij}(t|y)|\{t_l, f_{ij}^{(l)}\}_{l:y_l=y} \sim \mathcal{N}\left(m_{ij,y}(t), \Sigma_{ij,y}(t)\right) \tag{5.21}$$

with posterior mean function

$$m_{ij,y}(t) = \begin{bmatrix} m_{i,y}(t) \\ m_{j,y}(t) \end{bmatrix} \tag{5.22}$$

and posterior covariance kernel

$$\Sigma_{ij,y}(t) = \left(K_{ij,y}(t, t) + \sigma_n^2 I_\Delta\right) - K_{ij,y}(t, t_y)\left(K_{ij,y}(t_y, t_y) + \sigma_n^2 I_\Delta\right)^{-1} K_{ij,y}(t_y, t) \tag{5.23}$$

with matrix I_Δ given by

$$I_\Delta = \{\delta_{t_k, t_{k'}}\}_{k,k'}. \tag{5.24}$$

Through Bayes rule (5.12), we can obtain the log-odds corresponding to each of the $\binom{p}{2}$ trajectories

$$\phi_{ij}(t, f_{ij}) = \log \frac{\mathbb{P}\left(Y = +1 | \{t, f_{ij}\}, \mathcal{D}_{ij}\right)}{\mathbb{P}\left(Y = -1 | \{t, f_{ij}\}, \mathcal{D}_{ij}\right)} = \frac{1}{2} f_{ij}^T A_{ij}(t) f_{ij} + b_{ij}^T(t) f_{ij} + c_{ij}(t) \tag{5.25}$$

with parameters in the quadratic function given previously in (5.16) but replaced with their pair-wise values mean and covariance functions in (5.22) and (5.23). The training data for features i and j is given as $\mathcal{D}_{ij} = \mathcal{D}_i \cup \mathcal{D}_j$.

One may expand the existing dictionary of first order effects to incorporate additional second order effects by forming a new column vector ϕ, $\phi \leftarrow \{\phi_i(t, f_i)\}_{i=1}^p \cup \{\phi_{ij}(t, f_{ij})\}_{i,j>i}$ which is then inserted into (5.18) for extracting a single classifier. The addition of $\binom{p}{2}$ additional basis functions may be unrealistic in high-dimensional settings when p is large and $p > n$.

5.2.2 Kernel Function Parameter Estimation

The kernel covariance functions $k_{i,y}(t, t')$ contain parameters which must be estimated empirically. In this study, we will assume that the kernel covariance functions are of the form of the radial basis function. The kernel function of the i^{th} feature under class label y is given by

$$k_{i,y}(t, t') = \sigma_{i,y}^2 \exp\left(-\frac{\|t - t'\|^2}{2\ell_{i,y}^2}\right) \tag{5.26}$$

where $\sigma_{i,y}^2$ is the variance of the process for feature i under label y and $\ell_{i,y}^2$ is the characteristic length scale that determines how quickly the covariance between time points t and t' decays. Define $\theta_{i,y} = \{\sigma_{i,y}, \ell_{i,y}\}$ and $\theta = [\theta_1, \ldots, \theta_p, \sigma_n^2]^T$ with $\theta_i = [\theta_{i,+1}, \theta_{i,-1}]^T$. We seek to estimate θ via maximum likelihood of the trajectories under the GP prior distribution (5.5) with a zero mean prior function. For this estimation, the observed values are shifted to have zero mean by subtracting off their class

dependent mean, computed across all samples and time. The joint maximum log-likelihood problem is given as

$$\max_\theta \; l(\theta; \mathcal{D}) = \max_\theta \sum_{i=1}^{p} l(\theta_i, \sigma_n^2; \mathcal{D}_i) \tag{5.27}$$

with

$$l(\theta_i, \sigma_n^2; \mathcal{D}_i) = \sum_{y \in \{-1,+1\}} \sum_{l:y_l = y} \log \, p(f_i^{(l)} | t_l, y, \theta_{i,y}, \sigma_n^2) \tag{5.28}$$

and $p(f_i^{(l)} | t_l, y, \theta_{i,y}, \sigma_n^2)$ distributed by (5.5). We iteratively solve (5.27) via coordinate wise gradient descent [4] as the p marginal likelihoods are coupled by common noise parameter σ_n^2. Unfortunately, the log-likelihood in (5.27) is not convex and random restarts will be required to assess the global maxima [49, 4].

The partial derivative of (5.28) with respect to the l^{th} term in $\theta_{i,y}$, denoted by $\theta_{i,y,l}$, is given by

$$\frac{\partial}{\partial \theta_{i,y,l}} l(\theta_i, \sigma_n^2; \mathcal{D}_i) = \frac{1}{2} \sum_{j:y_j = y} \text{tr} \left((\alpha_{i,j} \alpha_{i,j}^T - K_{i,y_j}^{-1}(t_j, t_j)) \frac{\partial K_{i,y_j}(t_j, t_j)}{\partial \theta_{i,y,l}} \right) \tag{5.29}$$

with $\alpha_{i,j} = K_{i,y_j}^{-1}(t_j, t_j) f_i^{(j)}$. The partial derivative of (5.28) with respect to the common noise variance σ_n^2 is given by

$$\frac{\partial}{\partial \sigma_n^2} l(\theta_i, \sigma_n^2; \mathcal{D}_i) = \frac{1}{2} \sum_{y \in \{-1,+1\}} \sum_{j:y_j = y} \text{tr} \left((\alpha_{i,j} \alpha_{i,j}^T - K_{i,y_j}^{-1}(t_j, t_j)) \frac{\partial K_{i,y_j}(t_j, t_j)}{\partial \sigma_n^2} \right). \tag{5.30}$$

Note that if the time series are aligned, e.g., $t_j = t$, for all j, then we can avoid having to invert n_y different kernel matrices as $K_{i,y_j}(t_j, t_j) = K_{i,y_j}(t, t)$. The matrix derivatives of the kernel matrix with respect to the process variance, characteristic length scale, and noise variance for the radial basis function kernel (5.26) are given below

$$\frac{\partial K_{i,y_j}(t_j, t_j)}{\partial \sigma_i^2} = \left(\frac{1}{\sigma_i^2} K_{i,y_j}(t_{j_m}, t_{j_l}) \right)_{m,l} \tag{5.31}$$

$$\frac{\partial K_{i,y_j}(t_j,t_j)}{\partial \ell_i^2} = \left(\frac{\|t_{j_m} - t_{j_l}\|^2}{2(\ell_i^2)^2} K_{i,y_j}(t_{j_m}, t_{j_l}) \right)_{m,l}. \tag{5.32}$$

$$\frac{\partial K_{i,y_j}(t_j,t_j)}{\partial \sigma_n^2} = I. \tag{5.33}$$

Given $\hat{\theta}$, the kernel parameters for the GP trajectories are specified and the predictive posterior distributions within each log-odds basis function (5.13) can be formed. The structure of the resulting classifier will now be discussed using the loss function corresponding to the binomial deviance of logistic regression.

5.3 ℓ_1-Regularized Logistic Regression

When confronted an unlabeled test trajectory of p variables, we wish to place a prediction on the class that most likely produced such observations. In many modern disciplines, such as bioinformatics or signal processing, p tends to be large with $p \gg n$. In these high-dimensional settings, one must control the degrees of freedom of the resulting model. Additionally, many of the p features may not be capable of discriminating between different class labels, and thus, their influence in the final classifier should be removed in estimation. In this study, the penalty function $J(\beta)$ will be taken as the ℓ_1-norm, i.e., $\|\beta\|_{\ell_1} = \sum_i |\beta_i|$, which tends shrink many β_i's to zero for sufficiently large values of λ. We desire a value of λ such that $h(t, f)$ generalizes well to out of sample test data. This tuned value of λ, denoted by λ^*, is selected by minimizing the out-of-sample probability of error using 5-fold cross validation.

We will minimize the ℓ_1 regularized logistic regression loss function for forming the final classifier. The minimization is given by

$$\min_{\beta,\beta_0} \sum_{j=1}^{n} \log\left(1 + e^{-y_j\left(\beta^T \phi(t_j, f^{(j)}) + \beta_0\right)}\right) + \lambda \|\beta\|_{\ell_1}. \tag{5.34}$$

To solve (5.34), we use a gradient descent based method by quadratically expanding the loss function, resulting in iteratively solving a sequence of quadratic programs that incorporates an additional line search. At the m^{th} iteration of gradient descent, we seek the Newton step by solving the following

$$\delta\beta^{(m)} = \arg\min_{\beta} \frac{1}{2}\beta^T H^{(m)}\beta + \beta^T g^{(m)} + \lambda\|\beta\|_{\ell_1} \tag{5.35}$$

with gradient

$$g^{(m)} = \nabla_\beta L(\beta, \beta_0; \mathcal{D})|_{\beta_0=\beta_0^{(m)}, \beta=\beta^{(m)}} \tag{5.36}$$

and Hessian

$$H^{(m)} = \nabla_\beta^2 L(\beta, \beta_0; \mathcal{D})|_{\beta_0=\beta_0^{(m)}, \beta=\beta^{(m)}} \tag{5.37}$$

with logistic regression loss function $L(\beta, \beta_0; \mathcal{D}) = \sum_{j=1}^{n} \log\left(1 + e^{-y_j\left(\beta^T\phi(t, f^{(j)})+\beta_0\right)}\right)$ and updated parameter $\beta^{(m+1)}$ given by

$$\beta^{(m+1)} = \beta^{(m)} + \eta^{(m)}\delta\beta^{(m)} \tag{5.38}$$

with step size $\eta^{(m)}$ determined by performing a backtracking line search [4]. While (5.35) is convex, the presence of the ℓ_1-norm makes the objective function non-differentiable. However, the objective function can be transformed into an equivalent convex, differentiable objective by replacing the ℓ_1-norm with linear inequality constraints [4, 29] which produces a differentiable convex program in which any modern solver is capable of solving. The estimation of the kernel parameters, logistic regression coefficients, and the tuning parameter selection is summarized in Algorithm 3 (below).

Algorithm 3

1. Center all observed trajectories: $\{\tilde{f}_i^{(j)} \leftarrow f_i^{(j)} - \bar{f}_{i,y}\}_{i=1}^{p}, j \in \{j : y_j = y\}, y \in \{-1, +1\}$ where $\bar{f}_{i,y}$ is the scalar sample mean for feature i taken with respect to all samples from class y.

2. Obtain $\hat{\theta}$ via maximum likelihood (5.27) using gradient descent with random restarts on the centered Gaussian Process trajectories using $\tilde{\mathcal{D}} = \{t_j, \tilde{f}_i^{(j)}, y_j\}_{j=1,i=1}^{n,p}$.

3. ℓ_1-Logistic Regression

   (a) Initialize β, β_0

   (b) $\beta_0^{(m+1)} = \beta_0^{(m)} + \eta_0^{(m)} \delta\beta_0^{(m)}$ with $\eta_0^{(m)}$ via backtracking
   and $\delta\beta_0^{(m)} = -\left[\nabla^2_{\beta_0} L(\beta, \beta_0; \mathcal{D})\right]^{-1} \nabla_{\beta_0} L(\beta, \beta_0; \mathcal{D})|_{\beta_0=\beta_0^{(m)}, \beta=\beta^{(m)}}$

   (c) $\delta\beta^{(m)} = \arg\min_\beta \frac{1}{2}\beta^T H^{(m)} \beta + \beta^T g^{(m)} + \lambda\|\beta\|_{\ell_1}$

   (d) $\eta^{(m)} \leftarrow$ backtracking line search

   (e) $\beta^{(m+1)} = \beta^{(m)} + \eta^{(m)} \delta\beta^{(m)}$

   (f) Repeat steps 3b through 3e until some convergence criterion met

4. The optimal value of λ^* is chosen to minimize the out of sample probability of error using 5-fold cross validation, which indexes the final estimate of $\{\hat{\beta}(\lambda^*), \hat{\beta}_0(\lambda^*)\}$

5.4 Missing Time-Stamps in Test Data

We have presented an approach at extracting basis functions which summarize the discriminating power of multiple time-series observed at potentially mis-alligned time-points. The basis functions were obtained on a training set and used to form a single linear classifier in (5.34). Often times when handling out of sample test data, one may possess multiple observations forming a time-series but may be unaware of the corresponding time-stamps. This situation may occur when a patient makes

a sequence of visits to a physician in which data is gathered upon each visit. A schematic of an example involving the progression of a disease is shown in Figure 5.2. Here, t may represent the unknown time since infection, Δt_1 and Δt_2 represent the observable successive differences in time between subsequent visits to the physician. T is some maximum horizon time corresponding to the recovery period of the disease.

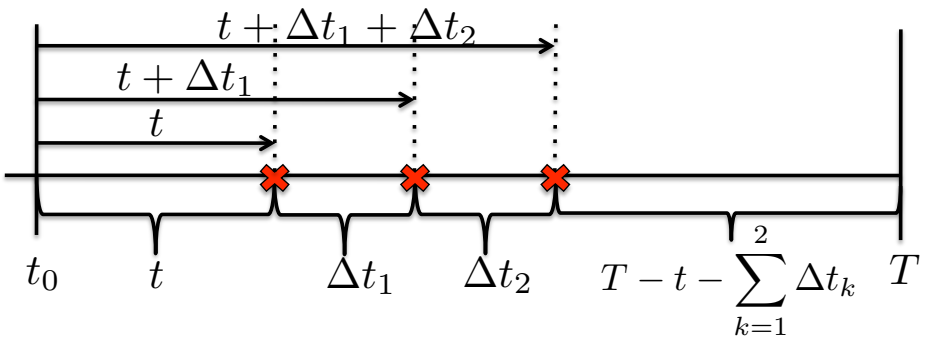


Figure 5.2: Sequential Measurements with Unknown Time Since "Infection" t

5.4.1 Unobserved Time as a Random Nuisance Parameter

Here, we will treat the missing time-stamps of an out of sample test trajectory as random nuisance parameters. We will assume that all kernel parameters, classification coefficients, and tuning parameters have been trained using the observed trajectories and their corresponding time-stamps. Adopting a prior distribution on t, denoted by $p(t|y)$, we can form the joint distribution of the sequence of observations for feature i and unobserved time stamps t conditioned on label y

$$p(f_i, t|y, \hat{\theta}_{i,y}) = p(f_i|t, y, \hat{\theta}_{i,y})p(t|y) \tag{5.39}$$

where $\hat{\theta}_{i,y}$ are the estimated kernel parameters and measurement variance for feature i and $p(f_i|t, y, \hat{\theta}_{i,y}, \hat{\sigma}_n^2)$ is the posterior distribution given in (5.9). We can marginalize

out the dependence of t to obtain the time-stamp independent posterior likelihood

$$p(f_i|y,\hat{\theta}_{i,y}) = \int_{t\in\mathcal{T}} p(f_i,t|y,\hat{\theta}_{i,y})dt = \int_{t\in\mathcal{T}} p(f_i|t,y,\hat{\theta}_{i,y})p(t|y)dt. \tag{5.40}$$

where $\mathcal{T}$ is the support of t. We can now define the time-independent log-odds

$$\phi_i(f_i) = \log \frac{p(Y=+1|f_i,\hat{\theta}_{i,y})}{p(Y=+1|f_i,\hat{\theta}_{i,y})} \tag{5.41}$$

which, when gathered over all p variables, is then inserted into the previously trained classifier in (5.17). The integral appearing in (5.40) may be computed using approximate Bayesian inference methods [19].

5.4.2 Unobserved Time as an Additional Class Label

Alternatively, one may desire a simultaneous prediction on both the label y and if t is within some region of time, e.g., phenotype = infected with Influenza A/H3N2 and time since inoculation is between 8 and 16 hrs. Here, we will decompose the time domain $\mathcal{T}$ into m, possibly unequal, time-regions

$$\mathcal{T} = \{\mathcal{T}_1 \cup \mathcal{T}_2 \cup \cdots \cup \mathcal{T}_m\},\ \mathcal{T}_i \cap \mathcal{T}_j = \emptyset,\ i \neq j. \tag{5.42}$$

Treating time as a random nuisance parameter, the joint likelihood of test trajectory f_i and $t \in \mathcal{T}_l$, is given by

$$p(f_i, t\in\mathcal{T}_l|y,\hat{\theta}_{i,y}) = \int_{t\notin\mathcal{T}_l} p(f_i,t|y,\hat{\theta}_{i,y})dt = \int_{t\notin\mathcal{T}_l} p(f_i|t,y,\hat{\theta}_{i,y})p(t|y)dt \tag{5.43}$$

which allows us to obtain the joint posterior probability of $\{Y=y, t\in\mathcal{T}_l\}$

$$p(Y=y, t\in\mathcal{T}_l|f_i,\hat{\theta}_{i,y}) = \frac{p(f_i,t\in\mathcal{T}_l|y,\hat{\theta}_{i,y})p(y)}{\sum_{y\in\{-1,+1\}}\sum_{l=1,\ldots,m} p(f_i,t\in\mathcal{T}_l|y,\hat{\theta}_{i,y})p(y)}. \tag{5.44}$$

One may then re-train the linear classifier in (5.17) to accommodate for multiple class labels using the basis functions resulting from (5.44).

5.5 Pan Viral Gene Expression Time-Series Results

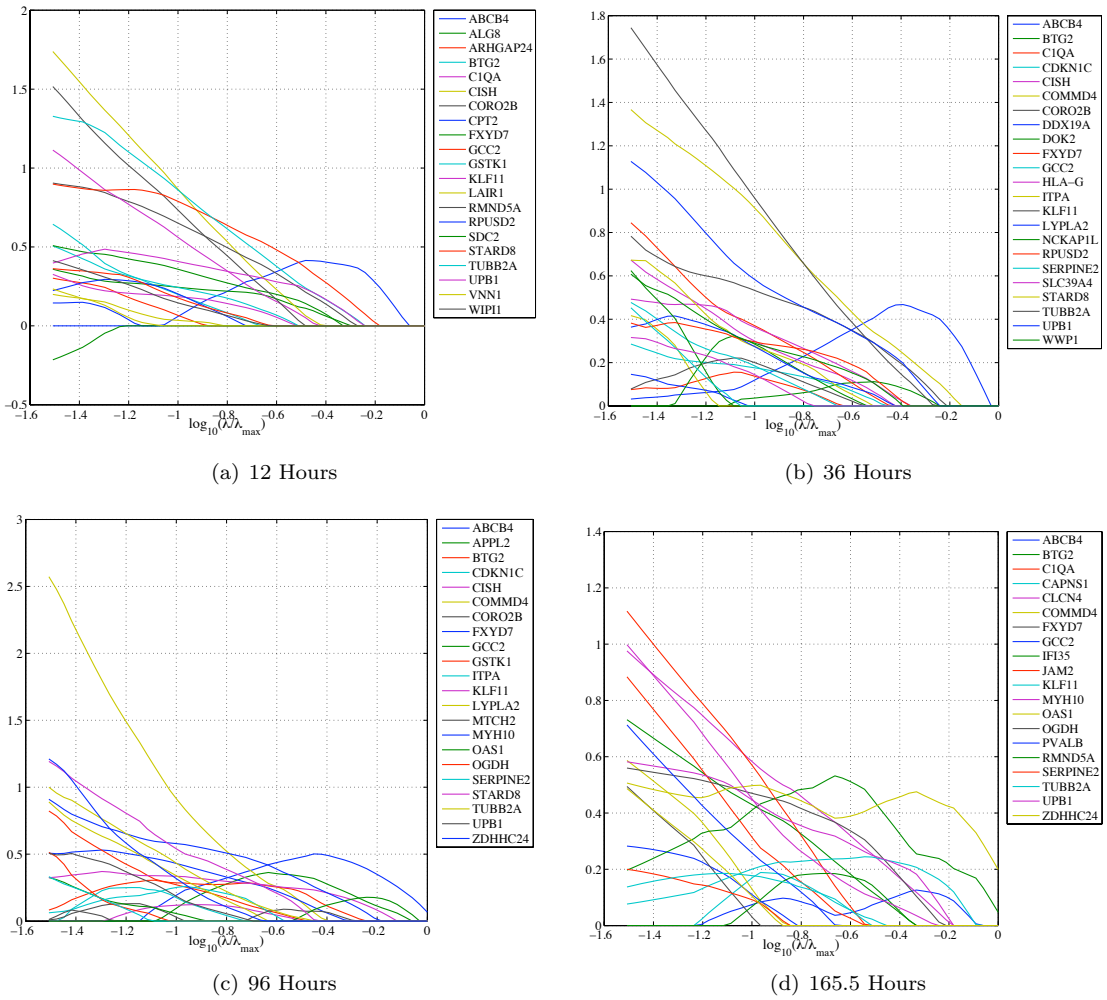
Here we present numerical results on peripheral blood time-series gene expression data set from a group of $n = 57$ patients inoculated with either HRV, H3N2, or RSV [67]. In a series of three challenge studies, patients were inoculated with one of the three viruses corresponding to that particular study. Roughly half of the patients responded with symptoms ($y = +1$, $n_{+1} = 28$) and the other half did not ($y = -1$, $n_{-1} = 29$). The original $12,023$ genes on the microarray were reduced to $p = 129$ differentially expressed genes controlling for a 20% False Discovery Rate [54]. The density of the sampling ranged from a minimum number of samples per patient of four to a maximum number of 20 with a median of 13. The time-series contain are not jointly aligned, although multiple patients have identical time-stamps corresponding to their peripheral blood samples. The goal is to explore the detection of symptomatic vs. asymptomatic using the proposed method as a function of increasing number of time-samples. We explore the ℓ_1-regularization paths and detection performance using LD and QD basis functions, resulting from assuming phenotype independent kernel parameters for the former and phenotype dependent kernel parameters for the latter.

We explored the effect of time-series duration on variable selection and detection performance by forming four sets of basis functions and corresponding linear classifiers using all samples with time-stamps up to and including 12 hrs, 36 hrs, 96 hrs and 165.5 hrs. The linear and quadratic discriminant basis functions are the log-odds ratio under posterior GP distributions. One would expect improved detection performance when including later time measurements as the symptomatic patients have fully developed the symptoms of their respective disease and the symptomatic pos-

terior mean functions should reflect this divergence from the asymptomatic posterior mean function and result in a non-zero basis log-odds ratio.

Figure 5.3 shows the ℓ_1-regularization paths using LD basis functions formed using all of the 57 patients samples up to and including 12 hrs, 36 hrs, 96 hrs, and 165.5 hrs corresponding to Figures 5.3(a), 5.3(b), 5.3(c), 5.3(d), respectively. LD kernel parameters were estimated using (5.27) over the 57 patients samples using 100 random restarts per gene-specific basis function and normalized to unit variance. The ℓ_1-regularization paths were obtained by inserting these LD basis functions into (5.34) and plotting $\hat{\beta}(\lambda)$ as a function of increasing λ. Inspection of the pre-symptomatic regime of up to and including 12 hrs post-incoluation in Figure 5.3(a), we see that genes ABCB4, TUBB2A, STARD8, and CORO2B are the first few genes which appear at large values of λ. Similar genes appear in Figure 5.3(b) at large values of λ. By 96 hrs, the patients have achieved peak symptoms. Figure 5.3(c) contains genes OAS1 and SERPINE2, both genes are typically activated in virus infection, which did not exist in the ℓ_1-regularization paths at the pre-symptomatic times of 12 and 36 hrs. By 165.5 hrs, OAS1 and IFI35 appear first at large values of λ. The appearance of virus activated genes at later time suggest that the GP basis functions, when presented with additional time points deeper into the perturbation study, are capturing the discrimination at peak symptoms over pre-symptomatic 12 hrs.

Figure 5.4 shows the ℓ_1-regularization paths using QD basis functions formed using all of the 57 patients samples up to and including 12 hrs, 36 hrs, 96 hrs, and 165.5 hrs corresponding to Figures 5.4(a), 5.4(b), 5.4(c), 5.4(d), respectively. QD kernel parameters were estimated by solving (5.27) over the 57 patients samples using 100 random restarts per gene-specific basis function. The ℓ_1-regularization paths were obtained by inserting these QD basis functions into (5.34) and plotting $\hat{\beta}(\lambda)$ as a



(a) 12 Hours

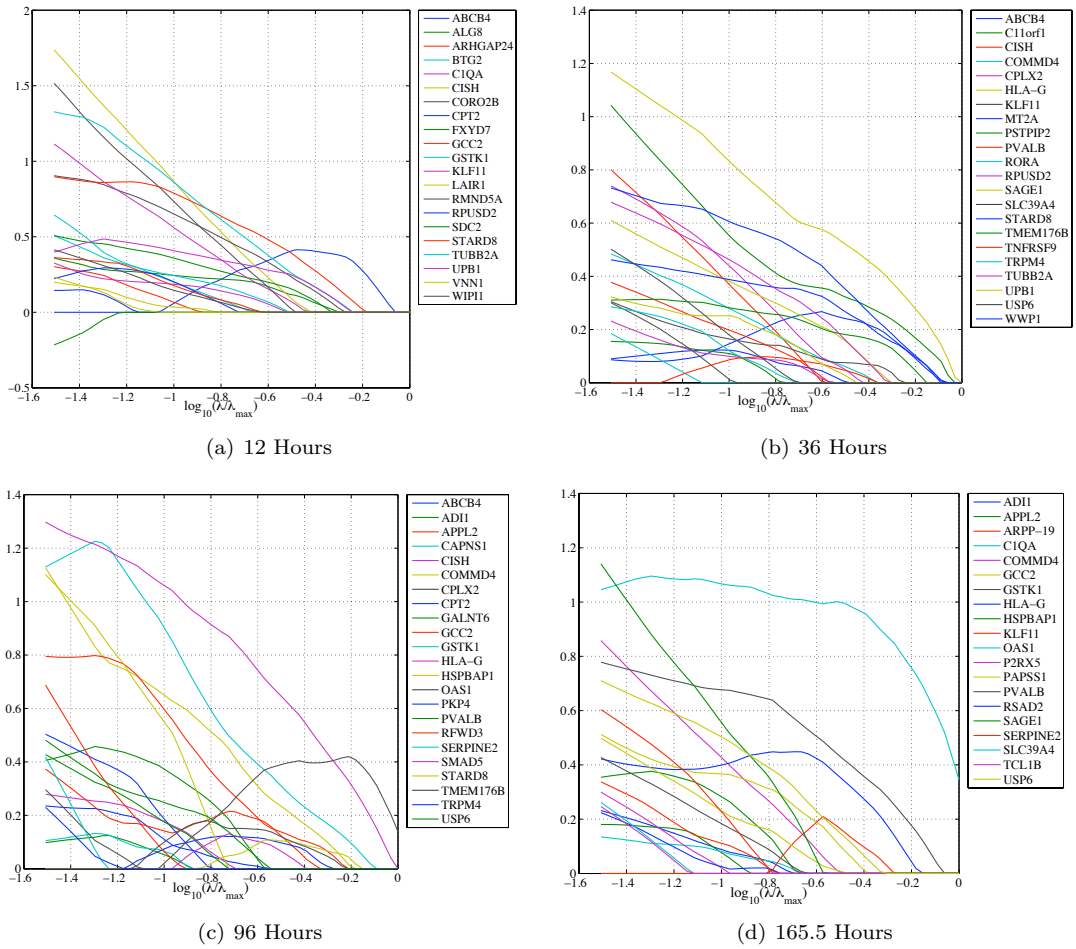
(b) 36 Hours

(c) 96 Hours

(d) 165.5 Hours

Figure 5.3: ℓ_1-Logistic Regression Regularization Paths Using Standardized Linear Discriminant Basis Functions Trained Using Subsets of Trajectories

function of increasing λ. Similarly to the ordering in Figure 5.3(a), we see in Figure 5.4(a), that genes ABCB4, TUBB2A, and STARD8 are the first few genes to appear at large values of λ. Genes OAS1 and SERPINE2 first appear in Figure 5.4(c) and corresponding to the earlier genes that appear in the regularization path. The anti-viral defense gene RSAD2 appears for the first time at 165.5 hrs, along with OAS1 at large values of λ. The appearance and disappearance of genes in the ℓ_1-regularization paths with quadratic discriminants further confirm that the functional nature of the GP basis functions are capturing time dependent discriminating behavior.



(a) 12 Hours

(b) 36 Hours

(c) 96 Hours

(d) 165.5 Hours

Figure 5.4: ℓ_1-Logistic Regression Regularization Paths Using Standardized Quadratic Discriminant Basis Functions Trained Using Subsets of Trajectories

The out of sample sensitivity, specificity, and probability of error were estimated with basis functions and final classifier trained using time-samples up to and including 12 hrs, 36 hrs, 96 hrs, and 165.5 hrs. For each of these four different time-horizon trained models, out of sample basis functions were formed using time samples up to and including 12 hrs, 24 hrs, 36 hrs, 48 hrs, 96 hrs, 117.5 hrs, and 165.5 hrs. These out of sample basis functions were formed using the posterior mean functions and covariance kernel functions in (5.13) estimated on the in-sample data corresponding to each of the four time horizons. To estimate the out of sample detection statistics, we held out a random subset of 11 patient's data. The remaining 46 p-dimensional

trajectories were used for training and validation. The tuning parameter λ was estimated by minimizing the 5 fold cross validated probability of error over these 46 p-dimensional trajectories. Given the optimal tuning parameter corresponding to the linear and quadratic discriminant basis functions (parameters estimated on this 46 patient in-sample), the final classifier models were formed using (5.34). Given the in-sample posterior mean functions, kernel functions, and $\hat{\beta}(\lambda^*)$, the sensitivity, specificity, and probability of error were assessed on this 11 patient out of sample data. This process was repeated 100 times for each of the four training time-horizons of 12 hrs, 36 hrs, 96 hrs, and 165.5 hrs. The trained models were applied to each of the seven out of sample test time-horizons.

Figure 5.5 shows the average out of sample (with error bars) probability of error, sensitivity, and 1-specificity trained using samples up to and including 12 hrs and applied to out of sample time samples up to and including 12 hrs, 24 hrs, 36 hrs, 48 hrs, 96 hrs, 117.5 hrs, and 165.5 hrs. By applying the classifier to samples with other time-horizons than the one the model was trained on, we can inspect the sensitivity of the classifier to out of sample test basis functions summarizing the divergence between phenotypes using data from additional or fewer time points. We see in Figure 5.5 a.) that the average probability of error at early times (less than 50 hrs) ranges from 0.31 to 0.35 using quadratic and linear discriminants, respectively. When the classifier with QDs is applied to data using 96 hrs, 117.5 hrs, and 165.5 hrs, the average probability of error is 0.35. However, when the classifier with linear discriminants is applied to these hours, the average probability of error decreases to 0.30. The sensitivity of the classifier using linear discriminants is less than that using quadratic discriminants. However, the classifier using QD produces an increase in 1-specificity as the out of sample time horizons increase whereas the classifier using

LD produces more stable 1-specificity.

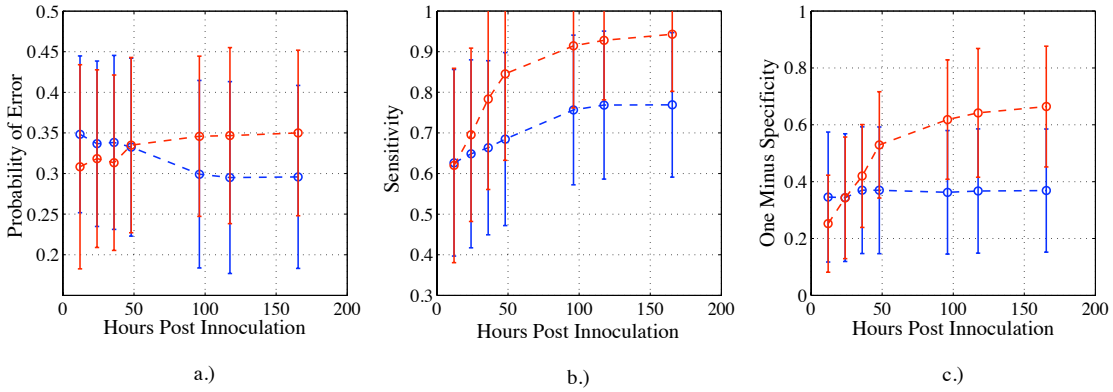


Figure 5.5: Probability of Error a.), Sensitivity b.), and One Minus Specificity c.) For ℓ_1-Logistic Regression Classifier Trained Using Samples up to and Including 12 Hours and Applied to Out of Sample Subsets of Trajectories up to and Including 12, 24, 36, 48, 96, and 165.5 Hours (Blue - Linear Discriminant, Red - Quadratic Discriminant)

Figure 5.6 shows the average out of sample (with error bars) probability of error, sensitivity, and one minus specificity trained using samples up to and including 36 hrs and applied to out of sample time samples up to and including 12 hrs, 24 hrs, 36 hrs, 48 hrs, 96 hrs, 117.5 hrs, and 165.5 hrs. Figure 5.6 a.) shows that the average probability of error at early times (less than 50 hrs) ranges from 0.34 to 0.27. When the classifier with quadratic discriminants is applied to data using 96 hrs, 117.5 hrs, and 165.5 hrs, the average probability of error for classifiers using both LDs and QDs is around 0.27. The classifiers using QDs over LDs appear to have slightly lower average probability of error. Both sensitivity and the 1-specificity increase with additional time-obsevations with the classifier using QDs having slightly higher values of both statistics.

Figure 5.7 shows the average out of sample (with error bars) probability of error, sensitivity, and one minus specificity trained using samples up to and including 96 hrs and applied to out of sample time samples up to and including 12 hrs, 24 hrs, 36

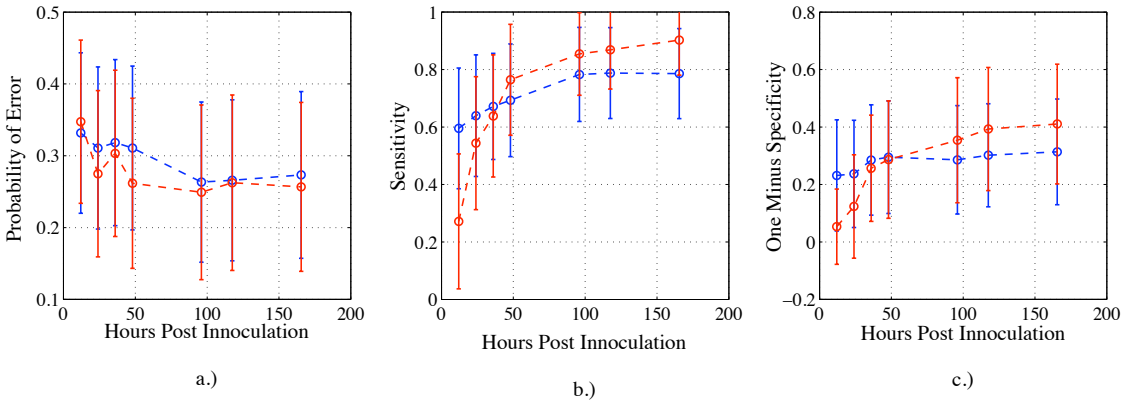


Figure 5.6: Probability of Error a.), Sensitivity b.), and One Minus Specificity c.) For ℓ_1-Logistic Regression Classifier Trained Using Samples up to and Including 36 Hours and Applied to Out of Sample Subsets of Trajectories up to and Including 12, 24, 36, 48, 96, and 165.5 Hours (Blue - Linear Discriminant, Red - Quadratic Discriminant)

hrs, 48 hrs, 96 hrs, 117.5 hrs, and 165.5 hrs. Figure 5.7 a.) shows that the average probability of error at early times (less than 50 hrs) decrease from 0.37 at 12 hrs to 0.32 at 50 hrs for both classifiers. When the classifiers are applied to data using 96 hrs, they both produce an average probability of error 0.23. The probability of error for the classifier using LDs decreases for 117.5 hrs and 165.5 hrs to 0.20 while the probability of error increases to 0.28 for these two time points using the classifier with QDs. The additional complexity of the model with quadratic basis functions may cause the poor generalization to these time points. The sensitivity for both methods increase with the time sample horizon with the LD based classifier producing higher average sensitivity. The 1-specificity for the classifier with LDs appears to be stable with increasing measurement time horizon whereas the classifier using QDs sees an increase in average 1-specificity.

Figure 5.8 shows the average out of sample (with error bars) probability of error, sensitivity, and one minus specificity trained using samples up to and including 165.5 hrs and applied to out of sample time samples up to and including 12 hrs, 24 hrs,

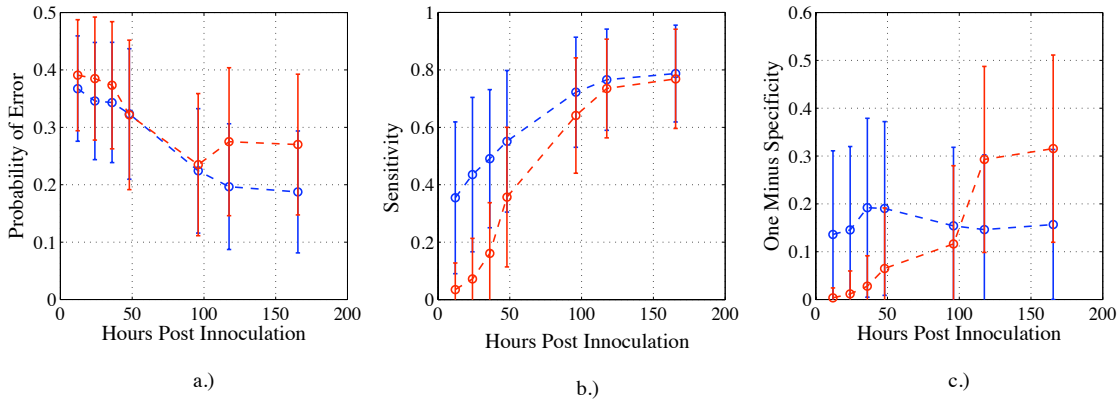


Figure 5.7: Probability of Error a.), Sensitivity b.), and One Minus Specificity c.) For ℓ_1-Logistic Regression Classifier Trained Using Samples up to and Including 96 Hours and Applied to Out of Sample Subsets of Trajectories up to and Including 12, 24, 36, 48, 96, and 165.5 Hours (Blue - Linear Discriminant, Red - Quadratic Discriminant)

36 hrs, 48 hrs, 96 hrs, 117.5 hrs, and 165.5 hrs. Figure 5.8 a.) shows that the average probability of error decreasing to 0.16 for both methods. The sensitivity for both methods increase with the time sample horizon with the LD based classifier producing higher average sensitivity. The 1-specificity for the classifier with LDs appears to be stable than the classifier using QDs.

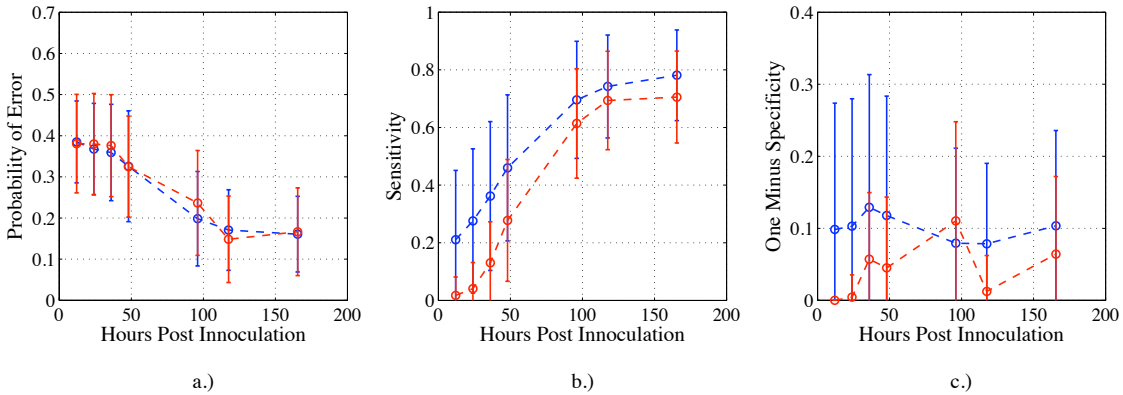


Figure 5.8: Probability of Error a.), Sensitivity b.), and One Minus Specificity c.) For ℓ_1-Logistic Regression Classifier Trained Using Samples up to and Including 165.5 Hours and Applied to Out of Sample Subsets of Trajectories up to and Including 12, 24, 36, 48, 96, and 165.5 Hours (Blue - Linear Discriminant, Red - Quadratic Discriminant)

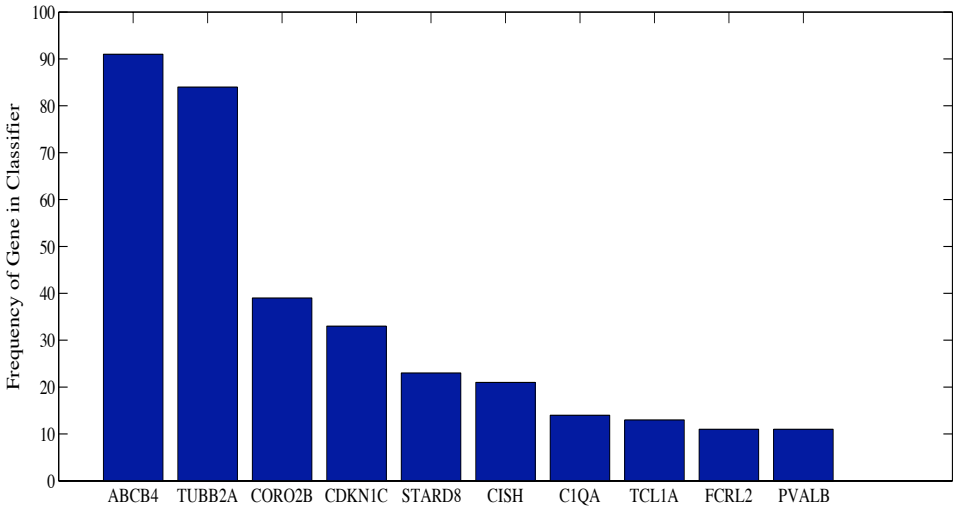
The results suggest that to accurately discriminate between phenotypes at early

times post-inoculation it is best to use a classifier trained using time samples in a neighborhood about these early time points. A classifier trained using early time samples performs well on trajectories at higher time points as the molecular divergence in the gene expression between phenotypes is strong. If one desired optimal classification performance above 96 hrs post inoculation, one should seek a classifier trained using full trajectories or in a neighborhood about these high times.

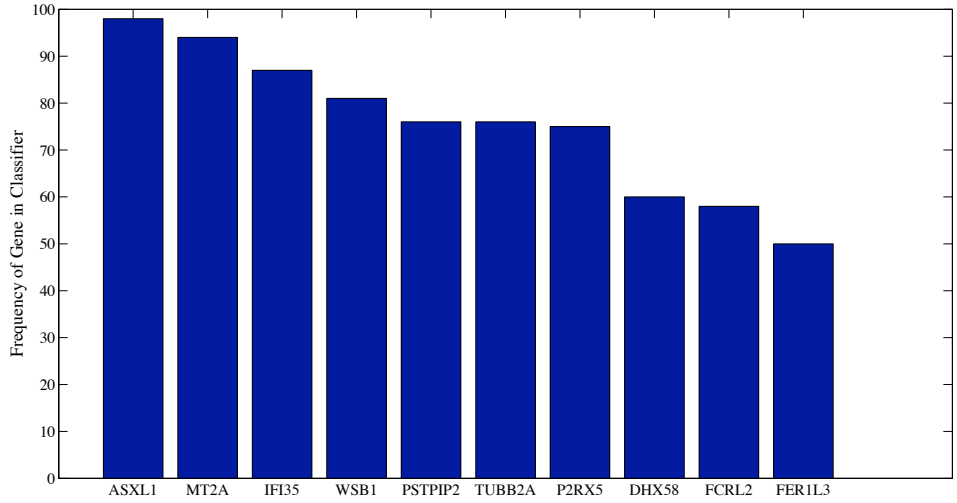
Figures 5.9, 5.10, 5.11, 5.12 depict the distribution of genes appearing in the final trained classifier within each of the 100 iterations of cross validation described above. The distribution frequency of genes appearing in the classifier with LDs tends to fall off sharply after the first few genes whereas the distribution corresponding to the classifier with QDs tends to have a more uniform distribution over differentially expressed genes. This suggests that the model with QDs is capturing higher order effects between the included genes than the classifier using LDs.

In many situations, one may be given a single static observation and corresponding time-stamp and must determine which phenotype generate such data. The proposed methodolgy naturally accommodates such situation given the model has been trained on full trajectories. Here, we explore the detection performance of the four classifiers trained using time samples up to and including 12 hrs, 36 hrs, 96 hrs, and 165.5 hrs and apply it to out of sample single observations from measurements gathered in neighborhoods about 12 hrs, 24 hrs, 36 hrs, 48 hrs, 96 hrs, 117.5 hrs, and 165.5 hrs. We define the neighborhoods for each of these seven times as the following: $\mathcal{T}_{12} = \{t : 8 \text{ hrs} \leq t \leq 16 \text{ hrs}\}$, $\mathcal{T}_{24} = \{t : 21.5 \text{ hrs} \leq t \leq 29 \text{ hrs}\}$, $\mathcal{T}_{36} = \{t : 30 \text{ hrs} \leq t \leq 42 \text{ hrs}\}$, $\mathcal{T}_{48} = \{t : 45.5 \text{ hrs} \leq t \leq 53 \text{ hrs}\}$, $\mathcal{T}_{96} = \{t : 84 \text{ hrs} \leq t \leq 101 \text{ hrs}\}$, $\mathcal{T}_{117.5} = \{t : 108 \text{ hrs} \leq t \leq 125 \text{ hrs}\}$, and $\mathcal{T}_{165.5} = \{t : 132 \text{ hrs} \leq t \leq 165.5 \text{ hrs}\}$.

Figures 5.13 shows the average detection performance (with error bars) of the



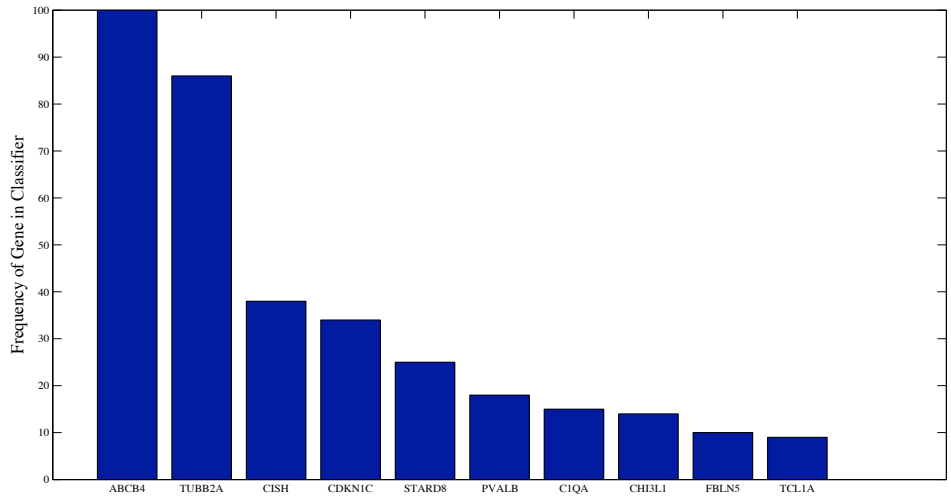
(a) Linear Discriminant Basis Functions



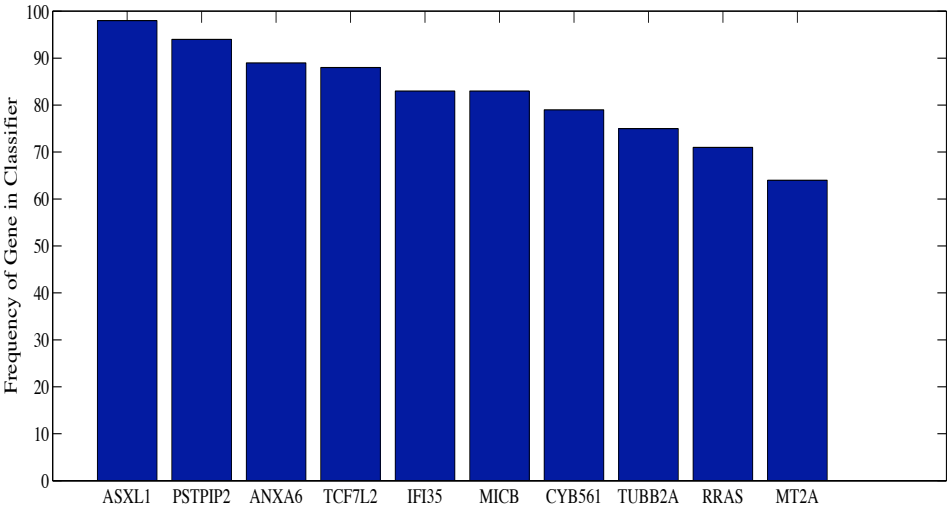
(b) Quadratic Discriminant Basis Functions

Figure 5.9: Distribution of Gene Frequency Appearing in "In-Sample" Trained Classifier Throughout the Cross Validation Trained on Samples Up to 12 Hours Post-Innoculation

classifiers trained using samples up to and including 12 hrs and applying it to single static measurements in neighborhoods about 12 hrs, 24 hrs, 36 hrs, 48 hrs, 96 hrs, 117.5 hrs, and 165.5 hrs. Both LD and QD based classifiers produce average probability of errors when applied to data around 12 hrs of 0.36. Both classifiers produce a decrease in average probability of error with the increase of samples deeper into the



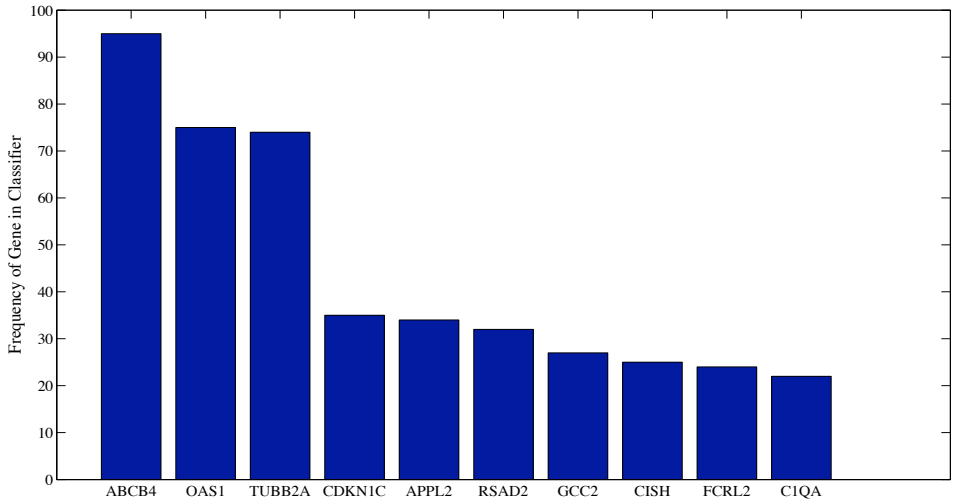
(a) Linear Discriminant Basis Functions



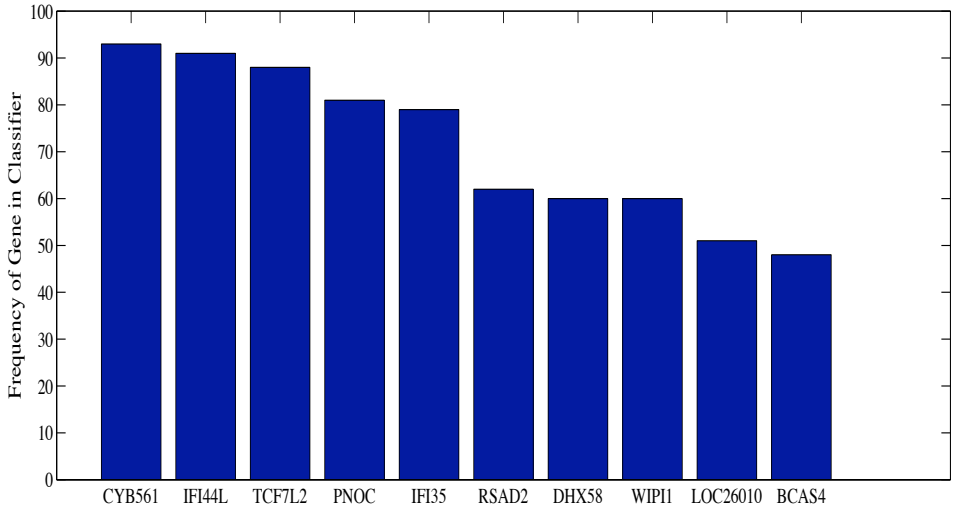
(b) Quadratic Discriminant Basis Functions

Figure 5.10: Distribution of Gene Frequency Appearing in "In-Sample" Trained Classifier Throughout the Cross Validation Trained on Samples Up to 36 Hours Post-Innoculation

perturbation. The QD based classifier produces average probability of error at the highest two time samples of 0.17 whereas the LD based classifier at these time points produces an average probability of error around 0.23. These results suggest that this QD based classifier, is capturing discriminating patterns better than the LD based classifier when applied to samples further into the perturbation experiment.



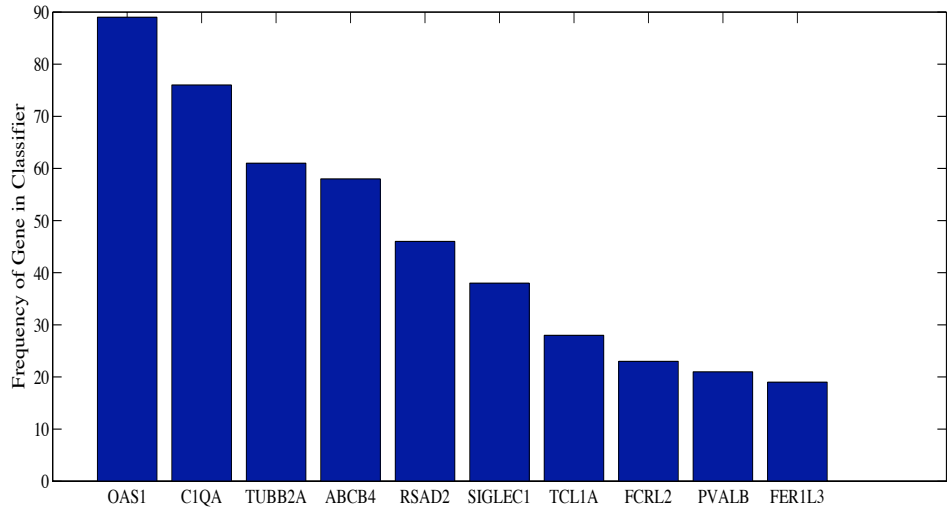
(a) Linear Discriminant Basis Functions



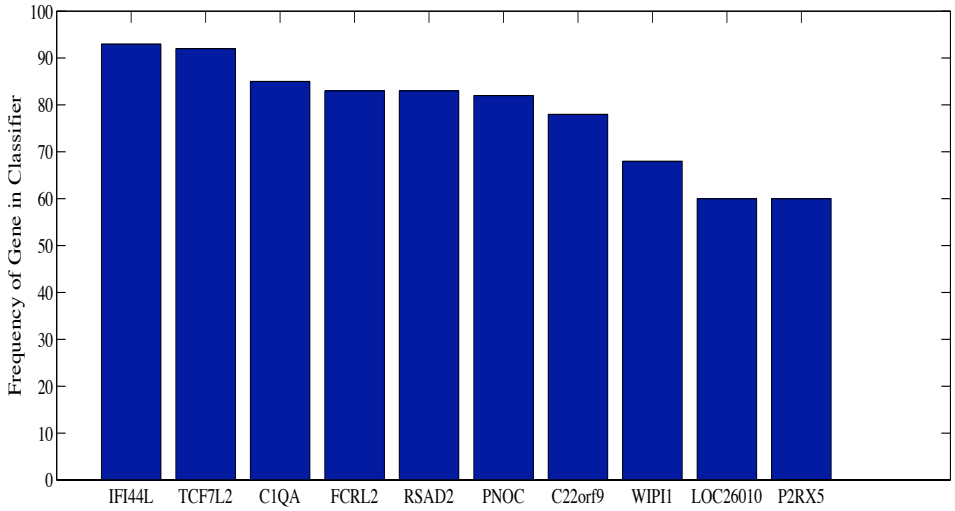
(b) Quadratic Discriminant Basis Functions

Figure 5.11: Distribution of Gene Frequency Appearing in "In-Sample" Trained Classifier Throughout the Cross Validation Trained on Samples Up to 96 Hours Post-Innoculation

Figure 5.14 shows the average detection performance (with error bars) of the classifier trained using samples up to and including 36 hrs and applying it to single static measurements in neighborhoods about 12 hrs, 24 hrs, 36 hrs, 48 hrs, 96 hrs, 117.5 hrs, and 165.5 hrs. Both LD and QD based classifiers produce average probability of errors when applied to data around 36 hrs of 0.30 and 0.34 for LD and QD, re-



(a) Linear Discriminant Basis Functions



(b) Quadratic Discriminant Basis Functions

Figure 5.12: Distribution of Gene Frequency Appearing in "In-Sample" Trained Classifier Throughout the Cross Validation Trained on Samples Up to 165.5 Hours Post-Innoculation

spectively. Both classifiers produce a decrease in average probability of error with the increase of samples deeper into the perturbation.

Figure 5.15 shows the average detection performance (with error bars) of the classifier trained using samples up to and including 96 hrs and applying it to single static measurements in neighborhoods about 12 hrs, 24 hrs, 36 hrs, 48 hrs, 96 hrs,

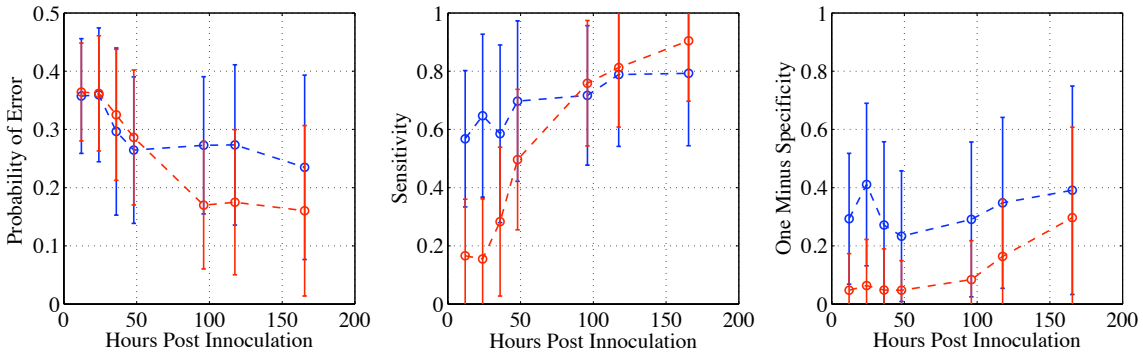


Figure 5.13: Probability of Error a.), Sensitivity b.), and One Minus Specificity c.) For ℓ_1-Logistic Regression Classifier Trained Using Samples at 12 Hours and Applied to Out of Sample Static Observations at 12, 24, 36, 48, 96, and 165.5 Hours (Blue - Linear Discriminant, Red - Quadratic Discriminant)

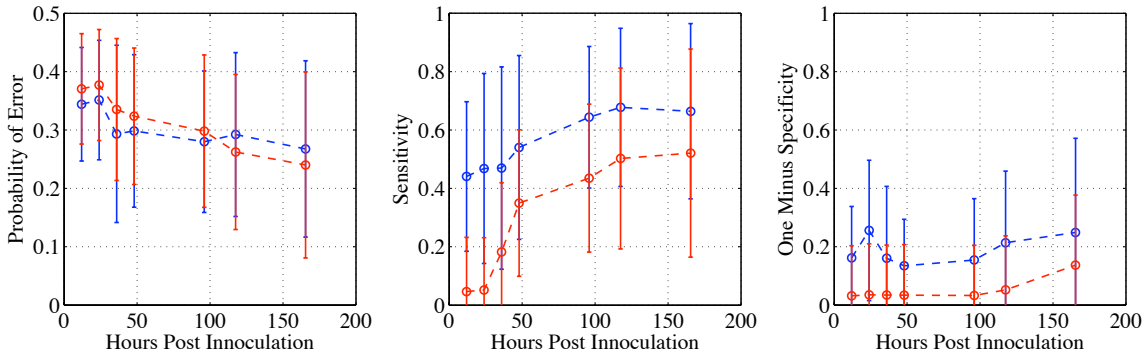


Figure 5.14: Probability of Error a.), Sensitivity b.), and One Minus Specificity c.) For ℓ_1-Logistic Regression Classifier Trained Using Samples at 36 Hours and Applied to Out of Sample Static Observations at 12, 24, 36, 48, 96, and 165.5 Hours (Blue - Linear Discriminant, Red - Quadratic Discriminant)

117.5 hrs, and 165.5 hrs. The classifier using LDs produces an average probability of error at 96 hrs of 0.23 whereas the classifier using QDs produces an average probability of error of 0.38. The results suggest that the increased complexity of the model, when using a single static observation, sees an erosion in its ability to generalize to out of sample data.

Figure 5.16 shows the average detection performance (with error bars) of the classifier trained using samples up to and including 165.5 hrs and applying it to single

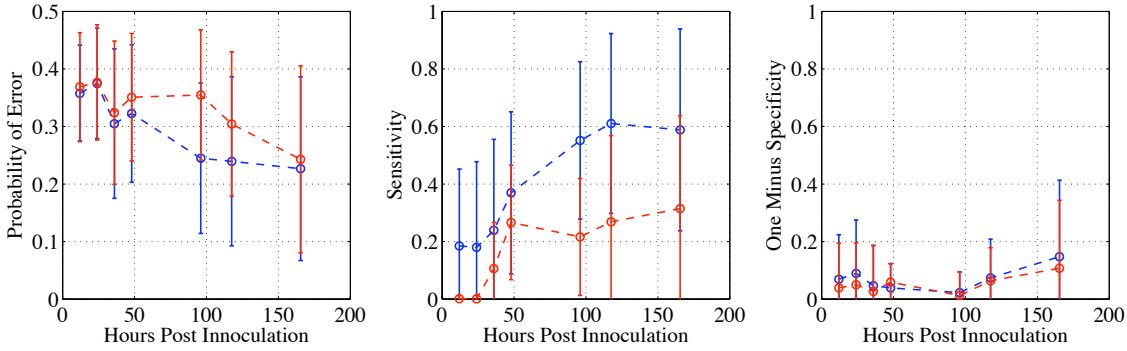


Figure 5.15: Probability of Error a.), Sensitivity b.), and One Minus Specificity c.) For ℓ_1-Logistic Regression Classifier Trained Using Samples at 96 Hours and Applied to Out of Sample Static Observations at 12, 24, 36, 48, 96, and 165.5 Hours (Blue - Linear Discriminant, Red - Quadratic Discriminant)

static measurements in neighborhoods about 12 hrs, 24 hrs, 36 hrs, 48 hrs, 96 hrs, 117.5 hrs, and 165.5 hrs. The classifier using LDs produces an average probability of error at 165.5 hrs of 0.23 whereas the classifier using QDs produces an average probability of error of 0.24. However, the classifier with LD basis function appears to generalize better to other time points than the classifier using QDs.

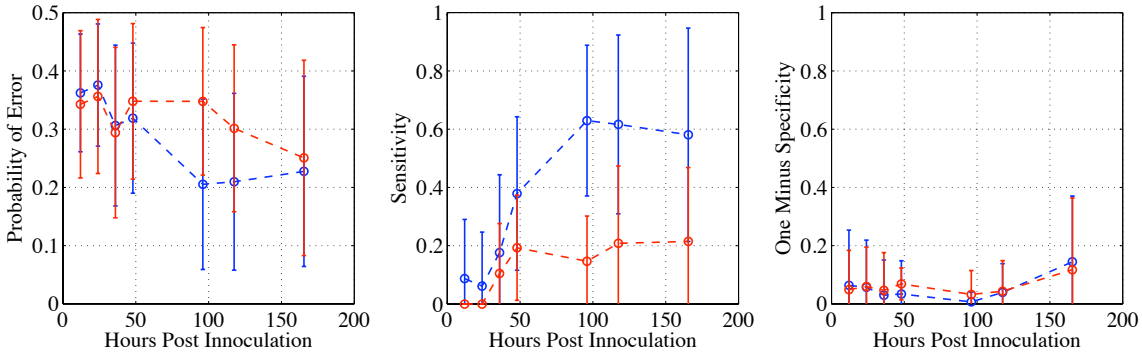


Figure 5.16: Probability of Error a.), Sensitivity b.), and One Minus Specificity c.) For ℓ_1-Logistic Regression Classifier Trained Using Samples at 165.5 Hours and Applied to Out of Sample Static Observations at 12, 24, 36, 48, 96, and 165.5 Hours (Blue - Linear Discriminant, Red - Quadratic Discriminant)

5.6 Conclusion

We have presented a methodology for the classification of multiple time-series by modeling the p-dimensional time-series are phenotype dependent Gaussian Processes. The discriminating power of the functional data is summarized as a log-odds ratio corresponding to each of the p-variables. The p-dimensional basis functions are fed into an ℓ_1-logistic regression classifier to generate a single powerful classifier. The proposed model can naturally accommodate mis-aligned time-series, a desirable feature when dealing with biological data. The method has been applied to a large pan viral human gene expression mis-aligned time-series data set and capable of accurately discriminating between asymptomatic and symptomatic patients at early and late times.

CHAPTER VI

Conclusion

The four core chapters of this dissertation are an attempt to develop necessary components of the PHD socio-molecular inference engine. The methods draw on insights and techniques from fields ranging from physics, epidemiology, statistical signal processing, operations research, bioinformatics, and machine learning. The concept of a spatio-temporal graphical model with high-dimensional noisy observations can model a social network with disease states changing over time and clinical observations being collected on the individual basis. Chapters 2 and 3 address the issue of reasoning under uncertainty on a large, potentially unknown social network, where limited resources are available for monitoring the state of the network. The latter two chapters confront the issue of accommodating noisy temporal individual high-dimensional data which would then be fused up into the social network level for performing inference on the hidden disease states.

While these four chapters develop the methodology for implementation of a PHD socio-molecular inference engine, additional work remains in extending these ideas to real time disease management. In particular, chapters 1 and 2 would greatly benefit from applying these methods to a controlled cohort of individuals where an infectious agent can propagate across a known social network. One could establish an estimate

on the social network topology given these observed disease trajectories and then insert this estimate into the adaptive sampling method. Statistical significance on the validity of the proposed adaptive sampling method and graphical model selection method could be quantified using a permutation test over an offline sampling strategy with an unknown estimate of the network topology. In such a cohort study, a patient's "state" is observed through measurement of gene expression, symptom history, etc. It would be worthwhile to use robust logistic regression to estimate this hidden state on a wild-type virus strain given a patients gene expression measurements using a model trained from existing challenge studies. This would quantify the extrapolative power of such a classifier and hopefully, outperform non-robust methods. Finally, one could establish the value added in gathering series of measurements from patients to better estimate their phenotype using the Gaussian Process framework discussed in chapter 4. We hope that these concepts can be successfully applied to such a challenge study and extended to active disease management.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] Leontine Alkema, Adrian E. Raftery, and Samuel J. Clark. Probabilistic projections of hiv prevalence using bayesian melding. *ANNALS OF APPLIED STATISTICS*, 1:229, 2007.

[2] Aharon Ben-Tal, Laurent E. Ghaoui, and Arkadi Nemirovski. *Robust Optimization (Princeton Series in Applied Mathematics)*. Princeton University Press, 2009.

[3] Dimitri P. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1999.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] R.J. Carroll, C.H. Spiegelman, K.K.G. Lan, K.T. Bailey, and R.D. Abbott. On errors-in-variables for binary regression models. *Biometrika*, 71(1):19–25, 1984.

[6] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. *ACM Trans. Inf. Syst. Secur.*, 10(4):1094–9224, 2008.

[7] S. Chandrasekaran, G. H. Golub, M. Gu, and A. H. Sayed. Parameter estimation in the presence of bounded data uncertainties. *SIAM J. Matrix Anal. Appl.*, 19(1):235–252, 1998.

[8] Edwin K. Chong, Christopher M. Kreucher, and Alfred O. Hero, III. Partially observable markov decision process approximations for adaptive sensing. *Discrete Event Dynamic Systems*, 19(3):377–422, 2009.

[9] Reuven Cohen, Shlomo Havlin, and Daniel ben Avraham. Efficient immunization strategies for computer networks and populations. *Physical Review Letters*, 91:247901, 2003.

[10] Gregory F. Cooper and Edward Herskovits. A bayesian method for the induction of probabilistic networks from data. *Mach. Learn.*, 9(4):309–347, 1992.

[11] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. *InterJournal*, Complex Systems:1695, 2006.

[12] A Doucet, S Godsill, and C Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3):197–208, JUL 2000.

[13] Moez Draief, Ayalvadi J. Ganesh, and Laurent Massoulié. Thresholds for virus spread on networks. *Ann. Appl. Probab.*, 18(2):359–378, 2008.

[14] Daniel Eaton and Kevin Murphy. Bayesian structure learning using dynamic programming and mcmc. In *UAI*, 2007.

[15] Laurent El Ghaoui, Gert R. G. Lanckriet, and Georges Natsoulis. Robust classification with interval data. Technical Report UCB/CSD-03-1279, EECS Department, University of California, Berkeley, Oct 2003.

[16] Stephen Eubank, Hasan Guclu, Anil, Madhav V. Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429(6988):180–184, 2004.

[17] J. Friedman, T. Hastie, and R. Tibshirani. A note on the group lasso and a sparse group lasso. arXiv:1001.0736v1, 2010.

[18] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostat*, 2007.

[19] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis, Second Edition*. Chapman & Hall/CRC, 2003.

[20] Laurent El Ghaoui and Hervé Lebret. Robust solutions to least-squares problems with uncertain data. *SIAM J. Matrix Anal. Appl.*, 18(4):1035–1064, 1997.

[21] Michael Grant, Stephen Boyd, and Yinyu Ye. CVX: Matlab Software for Disciplined Convex Programming. `http://www.stanford.edu/~boyd/cvx/`, Version 1.2, August 2008.

[22] T. Hastie, R. Tibshirani, and J. H. Friedman. *The Elements of Statistical Learning*. Springer, August 2001.

[23] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning, Second Edition: Data Mining, Inference, and Prediction*. Springer New York, 2009.

[24] A. O. Hero, D. Castenon, D. Cochran, and K. D. Kastella. *Foundations and applications of sensor management*. Springer, NY, 2007.

[25] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[26] Sabine Van Huffel, Ivan Markovsky, Richard J. Vaccaro, and Torsten Söderström. Total least squares and errors-in-variables modeling. *Signal Processing*, 2007, 2007.

[27] MJ Keeling and KTD Eames. Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295–307, SEP 22 2005.

[28] Seung-Jean Kim, Alessandro Magnani, and Stephen P. Boyd. Robust fisher discriminant analysis. In *In Advances in Neural Information Processing Systems*. MIT Press, 2006.

[29] Kwangmoo Koh, Seung-Jean Kim, and Stephen Boyd. An interior-point method for large-scale l1-regularized logistic regression. *J. Mach. Learn. Res.*, 8:1519–1555, 2007.

[30] Chris Kreucher, Keith Kastella, and Alfred O. Hero, III. Sensor management using an active sensing approach. *Signal Process.*, 85(3):607–624, 2005.

[31] Gert R. G. Lanckriet, Laurent E Ghaoui, Chiranjib Bhattacharyya, and Michael I. Jordan. A robust minimax approach to classification. *Journal of Machine Learning Research*, 3, 2002.

[32] K. Lange, D. R. Hunter, and I. Yang. Optimization transfer using surrogate objective functions. *Journal of Computational and Graphical Statistics*, 9, 2000.

[33] S.-I. Lee, V. Ganapathi, and D. Koller. Efficient structure learning of Markov networks using L1-regularization. In *Advances in Neural Information Processing Systems (NIPS 2006)*, 2007.

[34] Zhengdong Lu, Todd K. Leen, Yonghong Huang, and Deniz Erdogmus. A reproducing kernel hilbert space framework for pairwise time series distances. In *ICML*, pages 624–631, New York, NY, USA, 2008. ACM.

[35] David J. C. Mackay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, 2002.

[36] Lukas Meier, Sara van de Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal Of The Royal Statistical Society Series B*, 70(1):53–71, 2008.

[37] Nicolai Meinshausen and Peter Bühlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[38] Lauren Ancel Meyers. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bull. Amer. Math. Soc.*, 44:63–86, 2007.

[39] Kevin Patrick Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.

[40] M. E. J. Newman. The spread of epidemic disease on networks. *Physical Review E*, 66:016128, 2002.

[41] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[42] M E J. Newman. Threshold effects for two pathogens spreading on a network. *Physical Review Letters*, 95:108701, 2005.

[43] Brenda Ng, Leonid Peshkin, and Avi Pfeffer. Factored particles for scalable monitoring. In *In Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 370–377. Morgan Kaufmann, 2002.

[44] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2000.

[45] Mee Young Park and Trevor Hastie. L1 regularization path algorithm for generalized linear models. *J.R. Statist. Soc. B*, 69(4):659–677, 2007.

[46] Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

[47] Zinovi Rabinovich and Jeffrey S. Rosenschein. Extended markov tracking with an application to control. In *The Third International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 95–100, 2004.

[48] A.E. Raftery and L. Bao. Estimating and projecting trends in HIV/AIDS generalized epidemics using incremental mixture importance sampling. *Biometrics*, 2010.

[49] Carl E. Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[50] Pradeep Ravikumar, Garvesh Raskutti, Martin Wainwright, and Bin Yu. Model selection in gaussian graphical models: High-dimensional consistency of l1-regularized mle. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, and Léon Bottou, editors, *NIPS*, pages 1329–1336. MIT Press, 2008.

[51] Irina Rish, Mark Brodie, Sheng Ma, Natalia Odintsova, Alina Beygelzimer, Genady Grabarnik, and Karina Hernandez. Adaptive diagnosis in distributed systems. *IEEE Trans Neural Netw*, 16(5):1088–1109, 2005 Sep.

[52] Volker Roth and Bernd Fischer. The group-lasso for generalized linear models: uniqueness of solutions and efficient algorithms. In *ICML '08: Proceedings of the 25th international conference on Machine learning*, pages 848–855, 2008.

[53] M. Schmidt and K. Murphy. LassoOrderSearch: Learning Directed Graphical Model Structure using L1-Penalized Regression and Order Search. *Learning*, 8(34):2.

[54] A. Swaroop, A.J. Mears, G. Fleury, and A.O. Hero. Multicriteria Gene Screening for Analysis of Differential Expression with DNA Microarrays. *EURASIP Journal on Advances in Signal Processing*, 2004.

[55] R Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B-Methodological*, 58:267–288, 1996.

[56] R Tibshirani, M Saunders, R Rosset, and J Zhu. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, (67):91–108, 2005.

[57] Paul Tseng and Sangwoon Yun. A coordinate gradient descent method for nonsmooth separable minimization. *Math. Program.*, 117(1):387–423, 2008.

[58] R. H. Tütüncü, K. C. Toh, and M. J. Todd. Sdpt3 version 4.0 (beta) - a matlab software for semidefinite-quadratic-linear programming. `http://www.math.nus.edu.sg/mattohkc/sdpt3.html/`, July 2006.

[59] Martin J. Wainwright, Pradeep Ravikumar, and John D. Lafferty. High-dimensional graphical model selection using l1-regularized logistic regression. In *NIPS*, pages 1465–1472. MIT Press, 2006.

[60] A. Wiesel, Y.C. Eldar, and A. Beck. Maximum likelihood estimation in linear models with a Gaussian model matrix. *IEEE Signal Processing Letters*, 13(5):292, 2006.

[61] A. Wiesel, Y.C. Eldar, and S. Shamai. Optimization of the mimo compound capacity. *IEEE Transactions on Wireless Communications*, 6(3):1094, 2007.

[62] A. Wiesel, Y.C. Eldar, and A. Yeredor. Linear regression with Gaussian model uncertainty: Algorithms and bounds. *IEEE Transactions on Signal Processing*, 56(6):2194–2205, 2008.

[63] Huan Xu, Shie Mannor, and Constantine Caramanis. Robustness, risk, and regularization in support vector machines. *CoRR*, abs/0803.3490, 2008.

[64] Yang Yang, Jonathan D. Sugimoto, M. Elizabeth Halloran, Nicole E. Basta, Dennis L. Chao, Laura Matrajt, Gail Potter, Eben Kenah, and Ira M. Longini. The transmissibility and control of pandemic influenza a (h1n1) virus. *Science*, 326(5953):729–733, October 2009.

[65] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[66] Ming Yuan, Ming Yuan, Yi Lin, and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68:49–67, 2006.

[67] A.K. Zaas, M. Chen, J. Varkey, T. Veldman, A.O. Hero, J. Lucas, Y. Huang, R. Turner, A. Gilbert, R. Lambkin-Williams, et al. Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans. *Cell Host & Microbe*, 6(3):207–217, 2009.

[68] Alice X. Zheng, Irina Rish, and Alina Beygelzimer. Efficient test selection in active diagnosis via entropy approximation. In *Proceedings of the 21th Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, page 675, Arlington, Virginia, 2005. AUAI Press.

ABSTRACT

Inverse Problems in High-Dimensional Stochastic Systems Under Uncertainty

by

Patrick L. Harrington Jr.

Chair: Alfred O. Hero III.

Increasingly often, problems in modern medicine, quantitative finance, or social-networking involve tens of thousands of variables that interact with each other and jointly evolve over time. The states of these variables may correspond to the phenotype of a particular individual, the price of a security, or the current status of an individual's social networking profile. If these states are hidden to a researcher, additional information must be obtained to infer these hidden states based upon measurements of other variables, knowledge of the interacting network structure, and any dynamics that model the evolution of these states. This dissertation is an attempt to address general problems regarding reasoning under uncertainty in such spatio-temporal models but with an emphasis to applications in predictive health and disease in a loosely monitored population of individuals. The motivation is highly interdisciplinary and draws on tools and concepts from machine learning, statistics, epidemiology, bioinformatics, and physics.

We begin by presenting a solution to recursively sampling the best subset of nodes/variables that elicit the largest expected information gain of all sampled and un-sampled nodes in a large spatio-temporal complex network. We use methods from information theory and approximate Bayesian filtering to achieve this task. We then present a tractable method for empirically estimating the spatio-temporal graphical model structure corresponding to the "susceptible", "infected", and "recovered" (SIR) model of mathematical epidemiology. Here, we formulate the problem as an ℓ_1-penalized likelihood convex program and produce network detection performance superior to other comparable state of the art methods. We present a logistic regression classifier that is robust to worst-case bounded measurement uncertainty. The proposed method produces superior worst-case detection performance to the standard ℓ_1-logistic regression classifier on a Human rhinovirus (HRV) gene expression data set. The relationship between sparsity promoting regularization penalties and robustness to bounded measurement uncertainty is also established. The final chapter concludes with identifying the appropriate basis functions used in a classification model when the data is both high-dimensional and temporally sampled with ultimate goal of discriminating between multiple states/labels, e.g., phenotypes. We utilize Gaussian Processes and ℓ_1-logistic regression to accomplish this task and apply it to a human gene expression time-series data set resulting from a challenge study inoculation with Human Influenza A/H3N2, HRV, and Human respiratory syncytial virus (RSV).