

An Exploration of Data-Driven Techniques for  
Predicting Extreme Events in Intermittent  
Dynamical Systems

by

Stephen Carrol Guth

B.S, University of Maryland, College Park (2012)

Submitted to the Department Mechanical Engineering  
in partial fulfillment of the requirements for the degree of  
Master of ~~Engineering~~<sup>Science</sup> in Mechanical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2019

© Massachusetts Institute of Technology 2019. All rights reserved.


**Signature redacted**

Author .....

.....  
Department Mechanical Engineering  
July 16, 2019

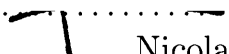
**Signature redacted**

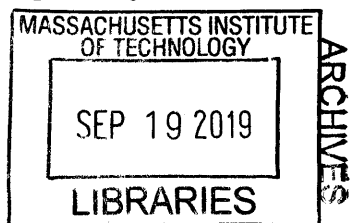
Certified by .....

.....  
  
Themistoklis P. Sapsis  
Professor of Mechanical and Ocean Engineering  
Thesis Supervisor

**Signature redacted**

Accepted by .....

.....  
  
Nicolas G. Hadjiconstantinou  
Chairman, Committee on Graduate Students





77 Massachusetts Avenue  
Cambridge, MA 02139  
<http://libraries.mit.edu/ask>

## **DISCLAIMER NOTICE**

Due to the condition of the original material, there are unavoidable flaws in this reproduction. We have made every effort possible to provide you with the best copy available.

Thank you.

**The images contained in this document are of the best quality available.**



# An Exploration of Data-Driven Techniques for Predicting Extreme Events in Intermittent Dynamical Systems

by

Stephen Carrol Guth

Submitted to the Department Mechanical Engineering  
on July 16, 2019, in partial fulfillment of the  
requirements for the degree of  
Master of Engineering in Mechanical Engineering

## Abstract

The ability to characterize and predict extreme events is a vital topic in fields ranging from finance to ocean engineering. Typically, the most-extreme events are also the most-rare, and it is this property that makes data collection and direct simulation challenging. In this thesis, I will develop a data-driven objective,  $\alpha$ -star, appropriate for optimizing extreme event predictor schemes. This objective is constructed from the same principles as Receiver Operating Characteristic Curves, and exhibits a geometric connection to scale separation. Additionally, I will demonstrate the application of  $\alpha$ -star to the advance prediction of intermittent extreme events in the Majda-McLaughlin-Tabak model of a dispersive fluid.

Thesis Supervisor: Themistoklis P. Sapsis

Title: Professor of Mechanical and Ocean Engineering





## Acknowledgments

I'd like to thank my adviser, Professor Themis Sapsis, for all the guidance, support, direction, and feedback he has provided over the last three years. It's no understatement to say I would never have crossed this finish line without his help and mentorship.

I'd also like to thank my adviser from my time at the United States Naval Academy (USNA), Professor Reza Malek-Madani, for setting me on this path at the beginning. Additionally from my time at USNA, I thank my 'co-advisors' Professors Kevin McIlhenny and Svetlana Abramov-Zamurovic as well as my other collaborators Professors Charles Nelson, Olga Korotkova, and Steve Wiggins. Without all of their assistance, I'd never have started the road to research.

My fellow researchers in the SAND lab deserve a special round of thanks for sharing their wisdom and expertise. Graduate students Zhong Yi Wan, Saviz Mowlavi, and Alexis Tzianis Charalampopoulos have each been vital beacons on the long road, whether it took the form of debating the mathematical models of ocean waves or the state of the art of Gaussian processes. At the same time, the SAND lab post-docs Mohammad Farazmand, Antoine Blanchard, and Hassan Arbabi have each been just as vital in orienting me in this field's special confluence of mathematics, engineering, physics, and computer science. I've had less time to share with the new grad student Rishabh Ishar, the recent graduates Han Kyul Joo and Mustafa Mohamad, and the post-doctorate ghost of Will Cousins, but they too have each helped me become the person able to write this thesis.

Lastly, this thesis would never have been completed without the generous support of the Office of Naval Research (ONR).



# Contents

<b>1</b>	<b>Introduction</b>	<b>17</b>
<b>2</b>	<b>Prior Literature</b>	<b>19</b>
2.1	Binary Classification . . . . .	19
2.1.1	Technique . . . . .	19
2.1.2	Evaluation . . . . .	20
2.1.3	Problems in the Extreme Event Context . . . . .	25
2.2	Dimensionality Reduction . . . . .	25
2.2.1	Main Theme . . . . .	25
2.2.2	Principal Component Analysis . . . . .	25
2.2.3	Diffusion Map . . . . .	27
2.3	Information Theoretic Quantities . . . . .	29
2.3.1	Preliminary Quantities . . . . .	29
2.3.2	Calculation from Data . . . . .	31
2.4	Causal Formulations . . . . .	33
2.4.1	Random Processes . . . . .	33
2.4.2	Shannon Entropy Rate . . . . .	34
2.4.3	Granger Causality . . . . .	34
2.4.4	Transfer Entropy . . . . .	35
2.4.5	Comments on Prediction . . . . .	36
2.5	Delay Embedding Theorems . . . . .	37
2.5.1	Overview . . . . .	37
2.5.2	Taken's Theorem . . . . .	37

2.5.3	Implementation Challenges . . . . .	38
2.5.4	Convergent Cross Mapping . . . . .	39
2.6	Mixed Methods . . . . .	40
2.6.1	Overview . . . . .	40
2.6.2	Ansatz Solution . . . . .	40
2.6.3	Variational Approach . . . . .	42
2.6.4	Closure Models . . . . .	43
2.6.5	Point of Departure . . . . .	43
2.7	Optimization Problems . . . . .	44
2.7.1	Overview . . . . .	44
2.7.2	Gradient Methods . . . . .	45
2.7.3	Interpolation Search . . . . .	46
2.8	Postmortem: Discarded Attempt . . . . .	48
2.8.1	Overview . . . . .	48
2.8.2	Retrospective Stencil . . . . .	48
2.8.3	Absolute Discrimination . . . . .	48
2.8.4	Interpretation . . . . .	49
2.8.5	Challenges . . . . .	50
<b>3</b>	<b>Receiver Operating Characteristic Metrics</b>	<b>51</b>
3.1	Problem Overview . . . . .	51
3.2	Groundwork . . . . .	52
3.2.1	Basic Quantities . . . . .	52
3.2.2	Basic Properties . . . . .	54
3.3	QRS Surface . . . . .	56
3.3.1	Precision Recall Curve . . . . .	56
3.3.2	Precision Recall Curve Metric . . . . .	57
3.3.3	Precision Recall Rate Surface . . . . .	58
3.3.4	Precision Recall Rate Surface Metric . . . . .	59
3.4	QRS Features . . . . .	60

3.4.1	Coinflip Predictor . . . . .	60
3.4.2	Knuckle . . . . .	60
3.4.3	Histogram Correspondance . . . . .	64
3.5	QRS Metrics . . . . .	67
3.6	Test Scenarios . . . . .	69
3.6.1	Bimodal Predictor . . . . .	69
3.6.2	Multivariate Gaussian Predictor . . . . .	71
3.6.3	Donut Predictor . . . . .	73
3.7	Appendix . . . . .	75
<b>4</b>	<b>Machine Learning Paradigm</b>	<b>77</b>
4.1	Problem Overview . . . . .	77
4.2	Hypothesis Class . . . . .	78
4.3	Training Data . . . . .	79
4.3.1	Overview of Issues . . . . .	79
4.3.2	Temporal Correspondence . . . . .	80
4.3.3	Spatial Correspondence . . . . .	81
4.3.4	Sampling Measure . . . . .	83
4.3.5	Summary . . . . .	84
4.4	Objective Function . . . . .	84
4.5	Optimization Algorithm . . . . .	85
4.5.1	Choice of Algorithm . . . . .	85
4.5.2	Parametrization of Search Space . . . . .	86
4.6	Potential Questions . . . . .	88
<b>5</b>	<b>Application I – Majda-McLaughlin-Tabak Model</b>	<b>89</b>
5.1	Model Overview . . . . .	89
5.2	Method . . . . .	92
5.2.1	Hypothesis Space . . . . .	92
5.2.2	Binning . . . . .	93
5.2.3	Objective Function . . . . .	93

5.2.4	Optimization Loop . . . . .	93
5.3	Results . . . . .	93
5.3.1	Features of Optimal Metric . . . . .	93
5.3.2	Comparison of Optimal Predictors . . . . .	94
5.3.3	Learning Rate . . . . .	97
5.3.4	Effects of Time Gap . . . . .	100
<b>6</b>	<b>Application II – Kolmogorov Flow Model</b>	<b>103</b>
6.1	Model Overview . . . . .	103
6.1.1	Method . . . . .	105
6.1.2	Results . . . . .	105
<b>7</b>	<b>Conclusions and Further Work</b>	<b>109</b>
7.1	Conclusion . . . . .	109
7.2	Further Work . . . . .	109

# List of Figures

2-1	Schematic diagram of the confusion matrix, representing false positives and false negatives (type I and II errors), as well as true positives and true negatives. . . . .	20
2-2	Sample Receiver Operating Characteristic Curve (ROC). . . . .	22
2-3	a) Plot of critical energy density $r_{\text{crit}}(L)$ in the MMT model, as a function of length scale $L$ . b) Plot of $r_{\text{crit}}(L, b)$ as a function of both length scale $L$ and background energy density $b$ . Figures taken from Cousins and Sapsis (2014) [67], Fig. 4 and 6. . . . .	41
3-1	Plot of a) a sample pdf and b) the corresponding precision recall (SR) curve. Each circle corresponds to a $(\hat{a}, \hat{b})$ pair: the curve is generated by fixing $\hat{a}$ and letting $\hat{b}$ vary. . . . .	56
3-2	Plot of a sample Precision Recall Extreme Event Rate (QRS) Surface. The surface is generated by varying $\hat{a}$ and $\hat{b}$ . . . . .	59
3-3	Plot of the QRS Surface for the coinflip predictor. $V = 0.5$ . . . . .	60
3-4	Precision-Rate slice of the QRS plot in figure 3-2, where $r = 0.5$ . The knuckle is captured as the non-decreasing interval just past $q = 0.1$ . . . . .	61
3-5	Sample PF plot when the second derivative $\frac{\partial^2 s}{\partial r^2}$ vanishes. . . . .	62
3-6	Sample PF plot when the second derivative $\frac{\partial^2 s}{\partial r^2}$ is a) always positive and b) always negative. . . . .	63
3-7	Sample PF plot when the second derivative $\frac{\partial^2 s}{\partial r^2}$ has one root. a) $\frac{\partial^2 s}{\partial r^2}$ is decreasing. b) $\frac{\partial^2 s}{\partial r^2}$ is increasing, and the first derivative is always positive. c) $\frac{\partial^2 s}{\partial r^2}$ is increasing, and the first derivative changes sign. . . . .	64



3-8	Sample pdf exhibiting scale separation. This demo is explored in section 3.6.1. . . . .	67
3-9	a) Joint pdf plots of the bimodal scenario for various parameters. b) Corresponding QRS plots. c, d) QRS metrics: c) $V$ , d) $\alpha^*$ . . . . .	70
3-10	Sample a) pdf plot and b) QRS surface for the multivariate Gaussian scenario. . . . .	71
3-11	Sample summary statistics for the multivariate Gaussian scenario. a) $V$ , b) $\alpha^*$ , c) $\eta_2$ , d) $\eta_1$ . . . . .	72
3-12	Sample a) pdf plot and b) QRS surface for the donut scenario, $R = 0.1$ . . . . .	73
3-13	Sample a) pdf plot and b) QRS plot for the nonmonotonically fixed donut scenario. $R = 0.1$ . . . . .	74
4-1	Cartoon representation of <b>strict time-lag</b> correspondence rule. . . . .	80
4-2	Cartoon representation of <b>space max</b> correspondence rule. . . . .	81
4-3	Sample learning curve for surrogate optimization. Note the change near $n = 20$ from pseudo-random samples to adaptive samples. . . . .	86
5-1	Sample plot of one simulated realization of the MMT model near an extreme event. . . . .	89
5-2	Probability density function of the MMR wave height. Rayleigh distribution overlaid for comparison—note the ‘long-tail’ extending from $x \approx 1.5$ to $x \approx 3.5$ . . . . .	90
5-3	Histogram of the number of extreme events per simulation run. . . . .	90
5-4	Cartoon representation of Gabor frame. . . . .	91
5-5	a) Sample prediction-truth joint pdf for a good MMT predictor. b) Corresponding ROC surface plot. . . . .	94
5-6	Optimal predictor parameters for each objective function. Note that total accuracy is very different than the others. . . . .	95
5-7	Receiver Operating Characteristic Curve comparisons of optimal predictors calculated via different objectives. a) precision-recall curve b) sensitivity-specificity curve. . . . .	95

5-8	Learning curves as a function of time. a) optimal parameters, b) parameter variance. . . . .	98
5-9	Learning curves as a function of data. a) optimal parameters, b) parameter variance. . . . .	99
5-10	a,b ) Optimal predictor parameters as a function of $\tau$ . c) Optimal $\alpha^*$ as a function of $\tau$ . . . . .	100
6-1	Descriptive plots for the Kolmogorov Flow. a) Sample realization of the the vorticity. b) Time series of energy dissipation near an extreme event. c) PDF of the energy dissipation . . . . .	103
6-2	Plots of single coefficient predictor quality for different wavenumbers and objectives. a) $\alpha^*$ , b) Volume under the surface, c) total accuracy, d) balanced accuracy, e) $F_1$ score. Note the consistent peak at (0, 4), which is resolved best by $\alpha^*$ and $F_1$ . . . . .	106
6-3	a) Composition of optimal predictor, in terms of Fourier modes, and b) quality of optimal predictor, each as a function of prediction gap $\tau$ . . . . .	106



# List of Tables



# Chapter 1

## Introduction

Many phenomena in a wide range of physical domains and engineering applications have observable properties that are normally distributed, that is, they obey Gaussian statistics. Gaussian-distributed random variables and processes are particularly easy to manipulate algebraically, and there is a rich literature using their properties in widely varying areas of probability and statistics, from Bayesian regression [13] to stochastic differential equations [28].

In some applications, however, random variables have significant non-Gaussian character. Frequently, the ‘long-tails’ of their distribution, which contain extreme, but rare events, are particularly important for a complete understanding of the phenomena in question. Examples of this behavior occur in ocean waves [21] and finance [34] [52], but similar phenomena can help explain behavior in fields as far afoot as cell dynamics [5] and mechanical part failure [68].

There are many approaches to modeling nonlinear (non-Gaussian) distributions, both parametric (bespoke distribution) and non-parametric (kernel methods). One well-studied method is the Gaussian Mixture Model, and that method provides approximate analytic values for various statistical and information theoretic quantities [1].

Unfortunately, while these approaches can capture the non-Gaussian nature of the distribution near the median, they don’t extend gracefully to the long tails or extreme events. A typical naive Monte Carlo approach to sampling a distribution

with a long tail will require samples proportional to the reciprocal of the probability of an extreme event occurring, which quickly becomes prohibitive for rare events.

More successful approaches to characterizing the long tail, such as large deviation theory [23] and probabilistic decomposition [42] instead take advantage of governing equations. In many physical situations, we can create a full set of equations describing the time evolution of a state vector, and leveraging these equations leads to effective solutions to questions about extreme event statistics and precursor [67].

In other situations, we have approximate models that require nonlinear copulae (‘closures’) in order to capture extreme event dynamics well. In these scenarios, a hybrid approach that combines equations and machine learning in order to determine the copulae can be successful [64].

Finally, in other situations, either our approximate linear models break down during extreme events, or we do not have approximate models at all. Instead, all we have are observations (‘Big Data’) and we want to characterize systems and in particular their extreme events in various ways. In thesis, I will discuss various mathematical tools appropriate for the question of identifying extreme events precursors from data. I will then construct a machine learning approach to find optimal precursors (‘predictors’), and show how to apply it to two model systems that display intermittent behavior.

# Chapter 2

## Prior Literature

### 2.1 Binary Classification

#### 2.1.1 Technique

##### Description

Binary classification is a supervised machine learning tool used in such fields as medical testing, information retrieval, and signal detection in which data points can be cleanly divided into two categories.

In supervised learning, data is provided that has been previously indicated as belonging to one of two categories, positive and negative. Some fraction of the data, the training set, is used to develop an algorithm which will predict which category a given data point will belong to. Then, the remainder of the data is used to validate the algorithm. The question arises, what metrics are best used to measure the goodness of the classification of the validation data?

A typical binary classifier will assign each data point a prediction score between 0 and 1. The data points scored closer to 1 are points the classifier believes most strongly to be positively scored. A threshold between 0 and 1 is introduced, based on desired features such as sensitivity or specificity. By varying the threshold, different balances between under-counting and over-counting true positives and negatives may be struck.



The literature for evaluating binary classification is voluminous. Kumari and Srivastava ([31]) have reviewed 21 studies sock puppet detecting, demonstrating over twenty different metrics for measuring the quality of binary classifications. Diettrich ([18]) has compared five statistical tests based on their ability to distinguish between classification algorithms.

## 2.1.2 Evaluation

### Basic Quantities

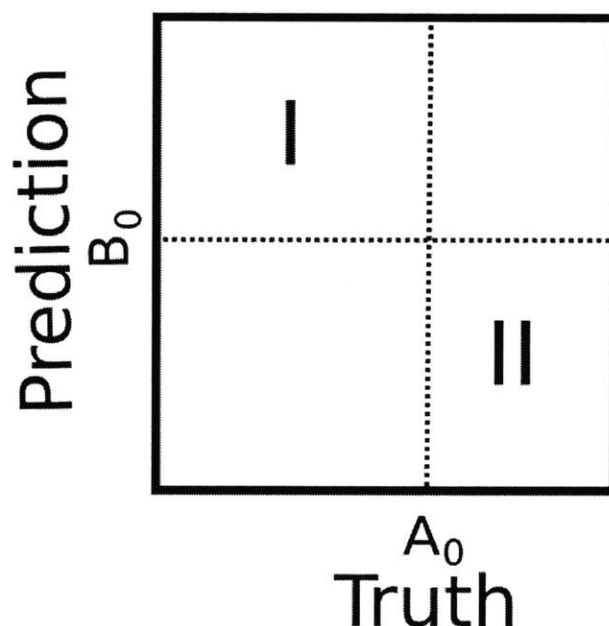


Figure 2-1: Schematic diagram of the confusion matrix, representing false positives and false negatives (type I and II errors), as well as true positives and true negatives.

A binary classification scheme produces four primitive quadrant counts, corresponding to

TP True Positives

FP False Positives

FN False Negatives

TN True Negatives

The basic evaluation metrics for a binary classification scheme are the True Positive Rate ( $TPR$ , also called Sensitivity) and the True Negative Rate ( $TNR$ , also called Specificity). These can be expressed as

$$\begin{aligned} TPR &= \frac{TP}{TP + FN} \\ TNR &= \frac{TN}{TN + FP}. \end{aligned}$$

As the binary classifier approaches perfect classification, both quantities approach 1.

Another way to characterize the basic evaluation metrics is Precision ( $q$ ) and Recall ( $r$ ), given by

$$\begin{aligned} q &= \frac{TP}{TP + FP} \\ r &= \frac{TP}{TP + FN}. \end{aligned} \tag{2.1}$$

Finally, an important characteristic of the underlying distribution is the Event Rate, given by

$$s = \frac{TP + FN}{TP + FN + FP + TN}.$$

Together,  $q$ ,  $r$ , and  $s$  (along with an overall scaling factor, the number of total counts  $n = TP + FP + FN + TN$ ) completely describe the binary classification and are enough to reconstruct the primitive counts.

### **Precision-Recall Trade-off**

In the medical diagnostic community, this trade-off is represented graphically in the Receiver Operating Characteristic Curve (ROC) [62]. The x-axis is chosen to be

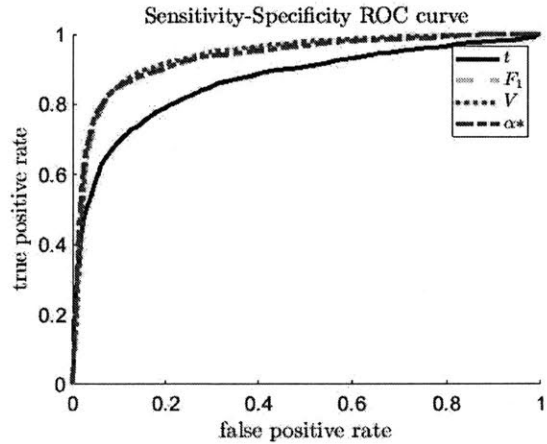


Figure 2-2: Sample Receiver Operating Characteristic Curve (ROC).

one minus the specificity, while the y-axis is chosen to be the sensitivity. The point  $(0, 1)$  then represents perfect classification and the diagonal line from  $(0, 0)$  to  $(1, 1)$  represents a completely random classifier.

A candidate classifier, applied to evaluation data, produces a classification with some sensitivity and specificity. Typically, the sensitivity may be varied, for instance by adjusting the pass threshold of the algorithm. The set of all such points (along with some joining rule) is the ROC.

One popular metric to evaluate the ROC is the Area Under the Curve (AUC, sometimes AUROC). Intuitively, a perfect classifier connects the points  $(0, 0)$ ,  $(0, 1)$ , and  $(1, 1)$  and has an integral of 1, while a random classifier connects  $(0, 0)$  to  $(1, 1)$  and has an integral of 0.5.

Another use of the ROC plot is to determine the best choice of threshold. In a review of threshold setting rules by Freeman and Moison ([19]), the authors noted that maximizing Youden's J statistic (sensitivity + specificity - 1) is equivalent to finding the point on the ROC curve where the tangent slope is equal to 1. Another threshold choice rule they compared was the ROC distance rule: choose the point on the ROC curve closest to the point  $(0, 1)$ .

One variant of the ROC relevant to rare event detection is the Concentrated ROC (CROC), developed by Swamidass et al. ([48]). The CROC is a nonlinear rescaling

of the ROC which emphasizes early retrieval, that is, the data points most certain to be classified as true positives.

Another variant is the Precision Recall Plot (PRC), which replaces the two axes of the ROC with precision and recall. Various reviews have shown that the PRC is better able to discriminate between classifiers when the data is wildly unbalanced ([27], [54]).

### **F-score**

One commonly used metric that balances between precision and recall is the F-score, given by

$$F_{\beta} = (1 + \beta^2) \frac{qr}{\beta^2q + r}, \quad (2.2)$$

where  $q$  is the precision and  $r$  is the recall. This is equivalent to a (weighted) harmonic mean, where the choice of  $\beta$  weighs more heavily precision (false positives) or recall (false negatives). Common choices for  $\beta$  include 0.5, 1, and 2.

One advantage of the F-score over other scalar metrics (such as Balanced Accuracy,  $B = \frac{1}{2}(TPR + TNR)$ ) is that equation 2.2 may be calculated without reference to the True Negative count. This is especially relevant in situations with unbalanced data (see subsection 2.1.2).

### **Algorithms**

Today, the most common approach to performing binary classification is called Support Vector Machine (SVM). SVM is a staple technique in Machine Learning, and its theory and implementation can be found in many standard textbooks [51].

In SVM, the algorithm seeks to find a hyperplane through the input space that best separates the data by maximizing a quantity called the geometric margin. When the data is not exactly separable, the size of the geometric margin trades between minimizing classification errors while maximizing classification confidence.

There is a tremendous literature on different implementations of SVM. The Pega-

algorithm in particular, is designed for in-line computation [50]. Further, Vapnik and Chervonenkis developed a rich theory of approximation error [61].

The objective function of SVM (usually log Hinge loss [37]) is usually augmented by a regularization term [29]. This helps to control for overfitting to noise, while also implicitly encoding a Bayesian-style prior.

Standard SVM is limited to a class of non-parametric linear hypotheses—the name-sake separating hyperplane. A standard extension to linear prediction problems is to augment the data vector with additional features: nonlinear functions of the other input data. This extension allows a linear algorithm to take advantage of particular nonlinear patterns in the data. Another extension, kernel-SVM, makes use of the kernel trick in order to construct nonlinear, non-parametric predictors using implicit nonlinear features.

One drawback of the non-parametric form is that physical interpretations of the SVM predictor functions are difficult. The SVM predictor is, in some sense, an explicit function of the training data—indeed, in the popular dual formulation of SVM the prediction is a linear combination of every data point in the margin (the “support vectors”).

## Unbalanced Data

In the extreme event context, one encounters the issue of unbalanced data. There are far fewer extreme events than non-extreme events, and the unbalanced training data sets cause problems for many of the more common techniques used. Solutions include balance independent metrics ([54]) and data resampling ([53]).

Sokolova and Lapalme [38] have shown that some but not all metrics may be invariant under certain changes in the confusion matrix, such as those that correspond to unbalanced data. In particular, they note that precision and recall (but not specificity/true negative rate) are invariant under the number of true negatives counts, an important property in the context of unbalanced data.

### 2.1.3 Problems in the Extreme Event Context

The major difficulty with posing the problem of extreme event prediction in the language of binary classification is reducing the nature of the extreme event to two categories. An extreme event is merely some event on the long tail of a probability distribution: there is rarely separation in the Rayleigh criterion sense. As a result, the question of threshold, not for predictor but for indicator, arises.

One way to bull through this issue is to collapse the distinction between ‘barely extreme’ events and ‘very extreme’ events. Another is to replace the binary classification scheme with a scheme that contains bins for many different classes, though many important algorithmic advantages are lost by the switch to multi-class classification ([56]).

## 2.2 Dimensionality Reduction

### 2.2.1 Main Theme

The application of dimensionality reduction techniques is as follows: suppose we have a large number of data points of the form  $(a_i, b_i)$ . If the shape of the distribution is very nearly one dimensional, then knowing  $b_i$  is very nearly ‘as good as’ knowing  $a_i$ . However, if the distribution cannot be reduced to one dimension without losing a great deal of information (quantified, perhaps, as variance), then knowing  $b_i$  alone is not enough to recover  $a_i$  accurately.

### 2.2.2 Principal Component Analysis

#### Algorithm

Principal Component Analysis (PCA) is a well-studied technique first developed by Pearson [46] in order to decompose a set of (linearly) correlated data points into uncorrelated variables. Each principal component has an associated variance, the fraction of the total variance explained away by that component. These components

can easily be used both for feature extraction and dimensionality reduction.

Given a zero-mean data matrix  $\mathbf{X}$  with covariance matrix  $\mathbf{X}^T\mathbf{X}$ , the transformation proceeds by identifying the loading vector  $w_1$  that maximizes its share of the data variance:

$$w_1 = \operatorname{argmax} \frac{w_1^T \mathbf{X}^T \mathbf{X} w_1}{w_1^T w_1}. \quad (2.3)$$

Put another way,  $w_1$  chosen so that after subtracting the  $w_1$  components, the data residuals have the least variance remaining.

This procedure is repeated until all of the data variance has been explained away, producing an orthogonal basis  $T = \{w_1, w_2, \dots, w_n\}$ . The variances associated with each basis vector are equal to the eigenvectors of the covariance matrix,  $\mathbf{X}^T\mathbf{X}$ .

### Application to Prediction

In two dimensions (one predictor, one indicator), PCA is a slight remapping of linear least squares regression.

### Limitations

PCA is a linear technique, with all the associated advantages and disadvantages. This assumption readily breaks down in both model and real nonlinear systems.

In particular, PCA assumes that the relationship between predictor and indicator is the same in all regions of the data space, even for extreme values. Even when the real relation can be effectively linearized, this linearization is often valid only for the ‘typical regime’ and it breaks down for extreme events.

Because the PCA algorithm works by diagonalizing the covariance matrix, it is sensitive to scaling issues. In particular, when different data axes have different physical units, there is no meaningful choice of relative scale. This is particularly acute in the case of constructed predictor functions, which may have *no* meaningful scale.

One common choice to diagonalize the correlation matrix rather than the covari-

ance matrix—in effect this choice normalizes for unit variance along each axis. See Leznik and Tofallis [41] for a more detailed description of this issue.

### 2.2.3 Diffusion Map

#### Overview

Diffusion Maps is a kernel-based nonlinear dimensionality reduction method developed by Coifman and Lafon [47]. It has been applied to spectral clustering [11], dynamical systems forecasting [58], and feature identification [59].

Diffusion Maps is motivated by the physical metaphor of heat diffusion, and using the rate of diffusion between two data points as a global distance metric.

#### Algorithm

The diffusion map procedure makes use of a heat kernel  $K$  given by

$$K(x, y) = \exp(-D(x, y)) \begin{cases} \exp(-D(x, y)) & D(x, y) \leq 3 \\ 0 & D(x, y) > 3 \end{cases}, \quad (2.4)$$

and a ‘bare Euclidean distance’ given by

$$D(x, y) = \frac{\|x - y\|}{\delta}, \quad (2.5)$$

where  $\delta$  is free parameter much larger than the typical inter-point spacing. The distance function has an artificial cutoff much larger than  $\delta$  for pairs for which  $K \ll 1$  (here set arbitrarily at  $3\delta$ ). This simplifies matrix math later, and is equivalent to a  $K$ -neighbors or  $\epsilon$ -ball decimation.

The diffusion kernel  $K$  is symmetric by construction. For the diffusion map procedure  $K$  is weighted by the following process



$$D(x) = \sum_y K(x, y) \quad (2.6)$$

$$L^{(\alpha)} = D^{-\alpha} K D^{-\alpha} \quad (2.7)$$

$$D^{(\alpha)} = \text{diag}(D) \quad (2.8)$$

$$M = D^{(\alpha)-1} L^{(\alpha)}. \quad (2.9)$$

When  $\alpha = 1$ , the procedure is ostensibly independent of the data sampling.

Find the eigenvalues and eigenvectors of the matrix  $M$ . The first eigenvalue ( $\lambda_1 = 1$ ) and eigenvector (uniform) are discarded. Then consider the next  $n$  largest eigenvalues and their associated eigenvectors  $\lambda_j$  and  $\phi_j$ .

Index the data points by  $i$ . Then  $j$ -th diffusion coordinate of the  $i$ -th data point is given by

$$X_{i,j} = \lambda_j^t \phi_j(i), \quad (2.10)$$

where  $t$  is a scale parameter. Increasing values of  $t$  cause the least important diffusion coordinates to contribute less and less.

### Limitations

Distributions with rare events typically feature *diffuse borders*, where the probability density drops away but only reaches zero after a great distance. In between, there is a region where  $p(\mathbf{x}) \approx N^{-1}$ , and data points are very spread out. The diffusion map kernel size suitable for the dense center of the distribution may be too small to reach these distant islands, and as a result they each ‘live’ in a separate diffusion coordinate.

One solution, proposed by Berry and Harlim ([60]) is to borrow the technique of variable bandwidth kernels. In this technique, the characteristic diffusion length is increased where the density of points is lowest. This has the advantage of eliminating spurious islands. It has the disadvantage of both reducing discrimination amongst

the extreme events, and introducing additional tuning parameters.

## 2.3 Information Theoretic Quantities

One way to talk about how well  $B$  predicts  $A$  is to ask how much information does an observation of  $B$  give us about  $A$ ? If each observation of  $B$  gives us no information about  $A$ , then  $B$  cannot possibly be a better predictor for  $A$  than mere guesswork. Conversely, if observations of  $B$  give a great deal of information about the state of  $A$ , we should be able to leverage that information to make good predictions.

### 2.3.1 Preliminary Quantities

#### Probability Distribution

For a continuous random variable  $A$ , we define the probability density function  $p(A)$  such that:

$$\begin{aligned} P(a_1 < A < a_2) &= \int_{a_1}^{a_2} p(A) dA \\ p(A) &\geq 0 \quad \forall A \\ \int p(A) dA &= 1. \end{aligned}$$

A fuller treatment of continuous distributions may be found in any elementary textbook on information theory, such as Cover and Thomas [57]. For our purposes, we will also need to define the joint distribution of two random variables, given by

$$p(A, B) = p(A) \cap p(B), \tag{2.11}$$

and the conditional distribution, given by

$$P(A|B) = \frac{p(A, B)}{p(B)}. \tag{2.12}$$

## Entropy

The entropy of a continuous distribution is given by

$$H(A) = - \int p(A) \log p(A) dA. \quad (2.13)$$

By a quirk of math, the entropy of a continuous distribution is not bounded below by zero. Nonetheless, distributional entropies can be compared:  $H(A) < H(B)$  can be interpreted as the statement that the distribution of  $A$  is narrower or more localized than the distribution of  $B$ .

This interpretation is easy to confirm in the case where  $\phi$  is a Gaussian random variable with covariance matrix  $\Sigma$ , in which case the entropy is given by

$$H(\phi) = \frac{1}{2} \log(2\pi e |\Sigma|^2). \quad (2.14)$$

where  $|\Sigma|$  is the determinant of the covariance matrix [32].

## Relative Entropy

The relative entropy, sometimes called the Kullback Leibler divergence, is given by

$$D_{\text{KL}}(A; B) = \int p(A) \log \frac{p(A)}{p(B)} dA.$$

## Mutual Information

The mutual information between two probability distributions is given by

$$I(A, B) = \int \int p(A, B) \log \frac{p(A, B)}{p(A)p(B)} dA dB, \quad (2.15)$$

or in terms of entropies as

$$I(A, B) = H(A) + H(B) - H(A, B) = H(A) - H(A|B). \quad (2.16)$$

The mutual information is the first quantity that might characterize the relationship between  $A$  and  $B$  for prediction purposes. Mutual Information is more general

than linear correlation [35], and in principal captures all of the relationship between  $A$  and  $B$  available from the data.

## 2.3.2 Calculation from Data

### General Notes on Calculation

In general, calculating mutual information from data involves all of the typical problems of deriving non-parametric distributions from data.

### Binning Method

One way to calculate these quantities is to replace integrals over continuous probability density functions with discrete sums over coarse-grained probability histograms. This replaces the problem of estimating the distribution with one of discretization.

Fraser and Swinney [22] describe the general partitioning technique: define a sequence of partitions  $G_0, G_1$  of the  $d$  dimensional space, where each partition consists of boxes. For each partition, the discrete mutual information can be expressed as

$$I^d(A^d, B^d) = \sum_j p(a_j^d, b_j^d) \log \frac{p(a_j^d, b_j^d)}{p(a_j^d)p(b_j^d)}. \quad (2.17)$$

At each step, one box ( $b_0$ ) is chosen according to some rule. Then, that box is partitioned into  $2^d$  sub-boxes ( $b_1^k, k \in [1, \dots, 2^d]$ ) according to another rule, typically one that attempts to match the counts of each sub-box.

If random variables along each axis are (approximately) independent over the box  $b_0$ , then the histogram counts of each sub box  $b_1^k$  should be (approximately) factorizeable. If this is the case, then box  $b_0$  needs no further partitioning. If, however, the histogram reveals deeper structure at the level of  $b_1^k$ , then define the next partition as the previous partition, with the refinement  $b_1^k$  in place of  $b_0$ . Continue with this process until for each box either there is no further substructure, or there are too few data counts for meaningful statistics on sub-boxes.

This procedure is readily generalizable to calculating entropy by simply replacing equation 2.17 with the expression for the entropy of a discrete distribution.

Partitioning techniques have a number of drawbacks, particularly related to dimensionality of the probability space. Adaptive schemes, such as the one developed by Darballay and Vajda [15], only partially address this limitation. Furthermore, Paninski [45] demonstrates particular convergence difficulties for finite data, and various bias correction schemes.

### Coarse Graining

The binning methods in the previous subsection are designed to accurately estimate the continuous entropy that would be achieved in the limit of infinitely small bins. As an alternate approach, one could replace the limiting or convergence process above with a target finite bin size.

Mutual information (and all these information theoretic quantities) is better behaved with discrete variables. Coarse graining is a commonly employed technique to deliberately decimate the data at some scale in order to take advantage of these simplifications. See, for instance, Katsoulakis and Plecháč (2013) [30] for an application of coarse grained information quantities in molecular systems.

Further, coarse graining is intimately tied to the physical thermodynamic interpretation of entropy, in which observable macrostates are composed of many unique but similar microstates.

### Nearest Neighbor Method

A different starting point for the estimation of mutual information is to make use of local topological properties of the data. Kraskov et al. [7] draw upon a rich literature of  $k$ -nearest neighbor methods for calculating the entropy of one dimensional distributions. The simplest entropy estimator is given by

$$H^d(X) = \frac{1}{N-1} \sum_{i=1}^{N-1} \log(x_{i+1} - x_i) + \psi(1) - \psi(N), \quad (2.18)$$

where  $x_{i+1} - x_i$  are the intervals between successive data points and  $\psi$  is the digamma function.

Kraskov et al.'s extension of this technique to the calculation Mutual Information is involved and attention is directed to their 2004 Physical Review E paper [7] for implementation details. In sketch, their method is based on comparisons between  $\epsilon$  balls in the joint space and each of the marginal spaces.

First, they interpret the definition of entropy (equation 2.13) as a sample space average of  $\log p(x_i)$ . This quantity is then expressed in terms of the number density of data points inside of a certain  $\epsilon$  ball centered at  $x_i$ . The size of this  $\epsilon$  ball is defined by the distance to the  $k$ -nearest neighbor, under the infinity norm. Finally, the authors judiciously choose different values of  $k$  for each of  $H(A)$ ,  $H(B)$ , and  $H(A, B)$  in order that certain statistical and systematic errors might approximately cancel when combined into  $I(A, B)$ .

The final estimator from Kraskov et al. [7] is given by

$$I^{(2)}(A, B) = \psi(k) - \frac{1}{k} + \psi(N) - \langle \psi(n_a) + \psi(n_b) \rangle, \quad (2.19)$$

where  $N$  is the number of data points,  $k \ll N$  is an algorithm choice, and  $n_a$  and  $n_b$  are the average data densities corresponding to the  $A$  and  $B$  marginal  $\epsilon$  balls.

Like all methods of calculating mutual information numerically from data, rigorous error statistics for non-Gaussian distributions are difficult to create. However, Kraskov et al. show better systematic error properties for multidimensional distributions using  $k$ -nearest neighbors compared to Darballay and Vajda's adaptive partitioning method [15].

## 2.4 Causal Formulations

### 2.4.1 Random Processes

To define the following quantities, we take  $A$  and  $B$  to be stationary zero-mean random processes.

Let  $A(t)$  be the value of the the random process  $A$  at time  $t$ . Then the  $p$  time lags of  $A$  may be expressed as  $A(t-1)$ ,  $A(t-2)$ , ...  $A(t-p)$ . Let  $A^{(p)}(t)$  be the  $1 \times p$

concatenated vector of lagged values, constructed as

$$A^{(p)}(t) = (A(t-1), A(t-2), \dots, A(t-p)). \quad (2.20)$$

Let  $A^-(t) = \lim_{p \rightarrow \infty} A^{(p)}(t)$  be the concatenated limit of all lagged values. Due to the stationarity assumption, explicit references to time  $t$  will hereafter be dropped.

### 2.4.2 Shannon Entropy Rate

The Shannon entropy rate for a discrete-time stochastic process is defined by

$$h = \lim_{n \rightarrow \infty} \frac{H(A^{(n+1)}) - H(A^{(n)})}{n} = \lim_{n \rightarrow \infty} \frac{H(A^{(n)})}{n}. \quad (2.21)$$

This equation may be justified from block entropy constructions, and various methods exist to use block entropies to estimate this quantity from sample processes ([8], [33]).

The entropy rate represents how much new information a source is generating over time. This is relevant to many information sciences areas, such as data compression. Here, the entropy rate serves as a link between quantities defined between distributions (such as Kullbeck-Leibler divergence and mutual information), and quantities defined for processes (such as transfer entropy and Granger Causality, discussed below).

### 2.4.3 Granger Causality

Granger causality was introduced by Wiener [66] and formalized by Granger [26] in the context of econometric models.

Let  $P_t(A|B^-)$  be the optimal predictor of  $A$  given access to the information contained in the past time lags of  $B$ , and let  $\sigma^2(A|B^-)$  be the variance of the residual errors.  $B$  Granger causes  $A$  iff  $\sigma^2(A|A^- \oplus B^-) < \sigma^2(A|A^-)$ . Put another way,  $B$  Granger causes  $A$  if past knowledge of  $B$  improves the prediction quality of  $A$  above and beyond past knowledge of  $A$  itself.

We borrow the final expression from Barnett et al. [36]: the Granger causality of  $B$  onto  $A$  is given by

$$F_{B \rightarrow A} = \log\left(\frac{\sigma^2(A|A^-)}{\sigma^2(A|A^- \oplus B^-)}\right), \quad (2.22)$$

and takes values on  $[0, \infty)$  that increase with causal strength.

In Granger's original paper, the author was careful to define all quantities relative to a background set  $D$  of all other relevant information in order to ferret out potential spurious third causes. In the context of prediction, this 'universe of potential causes' is restricted to the data available to potential predictors. In a practical scenario, it may well be the case that both buoy readings and wave heights are efficiently caused by some third quantity. Nonetheless, if the buoy readings Granger cause wave heights measurements then they satisfy the requirements for a good predictor.

There are two important approximations to be made to equation 2.22 in order to practically calculate it. First, instead of using the full time histories  $A^-$  and  $B^-$  as predictor inputs, the truncated lag sequences  $A^{(p)}$  and  $B^{(p)}$  are used instead. Second, instead of the optimal unbiased predictor  $P_t$  and associated residual variance  $\sigma^2$ , Granger suggested the use of the optimal linear predictor and its residuals.

Under these approximations, computation of the Granger Causality of  $B$  on  $A$  is reduced to a set of linear regression problems. While this is appropriate for many kinds of prediction, extreme events are frequently marked by a large degree of non-linearity.

#### 2.4.4 Transfer Entropy

Transfer entropy ([3]) is a quantity that measures the directed transfer of information from one process to another. It may be defined in terms of mutual information as

$$T(A, A^{(-)} \oplus B^{(-)}) = I(A, A^{(-)} \oplus B^{(-)}) - I(A, A^{(-)}) = I(A, B^{(-)}|A^{(-)}). \quad (2.23)$$



This quantity is motivated as the additional information about the transition probability of  $A$  that past knowledge of  $B$  provides above and beyond past knowledge of  $A$ .

For Gaussian processes, the entropy calculations in equation 2.23 can be simplified via covariance matrices (identity 2.23). Under this assumption, Barnett et al. showed that transfer entropy and Granger causality are equivalent [36].

Unfortunately, intermittent systems are rarely Gaussian and this correspondence will not apply to the prediction of rare events. Unlike Granger causality, transfer entropy inherits the invariance under nonlinear transformations characteristic of information theoretic quantities. This invariance may be compared to the practical decision of Granger to use optimal linear regression instead of the more general optimal predictor  $P_t$  (see, for instance, [35] for a comparison of mutual information with linear correlation).

#### 2.4.5 Comments on Prediction

All of these quantities have difficulties in the context of predicting extreme events that can be summarized as ‘nonconstructive.’

The Granger causality is defined in terms of an optimal predictor and its associated residual variance, but gives no suggestions on how to calculate it. Indeed, Granger’s suggestion is to search the class of linear predictors, and assume that the optimal linear predictor is good enough.

Transfer entropy calculations require numerical calculations of high dimension mutual informations, which is a daunting problem. But even this doesn’t quite solve the prediction problem. To say that the process  $B$  has a high transfer entropy into  $A$  is merely to say that knowing  $B$  should give the experimenter enough information to predict  $A$  well; it does not tell the experiment *how* to use that information to make predictions.

It is likely better to conceptualize these quantities as expensive ‘goodness measures’ for specific predictors, rather than as appropriate tools for constructing predictors.

## 2.5 Delay Embedding Theorems

### 2.5.1 Overview

In most practical dynamical systems, there is no direct access to measure the state of the system. Instead, there are a small number of observable functions, the measurement of which produces a set of time series. This restriction is intuitively sensible: rather than measure Euclidean velocity in the ocean at every point at all times, we only have access to measurements at a finite number of buoys (or, depending on satellite coverage, access to a certain set of measurement restricted to the sea surface).

A *delay embedding theorem* gives the conditions under which the time series of the observable may be used to recover the state of the underlying dynamical system.

### 2.5.2 Taken's Theorem

Floris Takens (1981) [55] proved that, assuming no special symmetries, a sequence of  $2m + 1$  delays would always suffice to reconstruct an underlying compact manifold of dimension  $m$ . Formally, the theorem states:

**Theorem 1.** *Let  $M$  be a compact manifold of dimension  $m$ . For pairs  $(\phi, y), \phi : M \rightarrow M$  a smooth diffeomorphism and  $y : M \rightarrow \mathcal{R}$  a smooth function, it is a generic property that the map  $\Phi_{\phi, y} : M \rightarrow \mathcal{R}^{2m+1}$ , defined by*

$$\Phi_{\phi, y}(\mathbf{x}) = (y(\mathbf{x}), y(\phi(\mathbf{x})), \dots, y(\phi^{2m}(\mathbf{x}))), \quad (2.24)$$

*is an embedding; by "smooth" we mean at least  $C^2$ .*

Restated in less technical jargon,  $\phi$  defines the time evolution of some discrete-time dynamical system with state  $\mathbf{x}$ . That is to say, given  $\mathbf{x}(t)$ , we have  $\phi(\mathbf{x}(t)) = \mathbf{x}(t + \tau)$

Whatever the dimension of the space that  $\mathbf{x}$  is defined on, the dynamics are restricted to some  $m$ -dimensional manifold  $M$ . While we don't have access to the true state  $\mathbf{x}$ , we do have access to some scalar observable function  $y(\mathbf{x})$ .

By measuring the time series of  $y(\mathbf{x})$ , we can construct a sequence of time lags

$\Phi_{\phi,y}(\mathbf{x}) = (y(\mathbf{x}), y(\phi(\mathbf{x})), \dots, y(\phi^n(\mathbf{x})))$ . This sequence is an embedding of the dynamical manifold  $M$ ; that is, all of the interesting features of  $M$  are preserved in  $\Phi_{\phi,y}(\mathbf{x})$ , and any calculations depending on  $M$  should be calculable with just  $\Phi_{\phi,y}(\mathbf{x})$ .

Takens's result was proved in the context of strange attractors, where there is research interest in the dimension and entropy of the attracting surface of some chaotic system. See, for instance, Arbarbanel (1996) [4] for an in-depth description of how to apply this type of delay embedding theory to study chaotic systems.

### 2.5.3 Implementation Challenges

There are a few difficulties in translating this theory into practical applications in systems featuring extreme event:

- ODE vs PDE systems
- sensitivity to measurement error
- high dimensional computation concerns

First, the theorem is proven for dynamical systems on a compact manifold of dimension  $m$ . That is to say, a system of *ordinary* differential equations of finite dimension. Many of systems of interest in extreme events are best defined as *partial* differential equations. While there are standard methods of reducing PDE systems to ODE systems (finite difference, finite volume, finite element, etc), the immutability of a single dimension  $m$  is mortally weakened.

This is doubly so in the case of chaotic PDE systems, which are marked by a continuous spectra that corresponds to an infinite dimensional manifold. The best that can be hoped for is some finite truncation of the  $m$  most significant eigen-modes.

Second, while Takens's theorem demands that the observable  $y$  be smooth, it makes little claim about the smoothness and sensitivity of the embedding  $\Phi_{\phi,y}(\mathbf{x})$ . This is a problem, because any real measurement of  $y(\mathbf{x})$  will be subject to some measurement error  $\Delta y$ . If the geometry of the recovered embedding depends on precise relative measurement of two quantities measured some long period apart in

time ( $2m\tau$ ), the effective measurement error may grow like  $\exp(2m\tau\lambda)$ , where  $\lambda$  is the dynamical system's greatest Lyapunov exponent.

Attempts to ameliorate this trend, by taking short time delays so that  $\tau \ll \frac{1}{\lambda}$ , run into an opposite sort of numerical instability: a set of different vectors  $\Phi_{\phi,y}(\mathbf{x})$  (for different initial  $\mathbf{x}_k$ , for instance) will be very nearly collinear in the vector space  $\mathcal{R}^{2m+1}$ . Put equivalently, there is implicitly a linear transformation  $A$ , and the matrix  $A$  is very badly conditioned. Thus, small errors in the measurement of  $y(\mathbf{x})$  again propagate and become large errors in the embedding  $\Phi_{\phi,y}(\mathbf{x})$ .

Third, one immediate use of the recovered embedding is in calculating distributional entropies. While entropy is a well defined quantity in high dimensional contexts (see, for instance, Lesne 2014 [33]), numerical calculations quickly become intractable as the dimension increases.

Unfortunately, systems exhibiting extreme events are invariably high dimensional, and it is not clear that the tricks used for dimensionality reduction ([42], [44]) translate into this approach.

## 2.5.4 Convergent Cross Mapping

In order to quantify causality in weakly coupled dynamical systems, Sugihara et al. ([25]) developed an alternative to Granger Causality called Convergent Cross Mapping (CCM). This technique compares lagged time series drawn from different scalar observables to find cross correlations from one to the other. In particular, a certain difference in directed correlations leads to the inference that one observable is causally correlated with the other (though see rebuttals, [40] and [14]).

The machinery underlying CCM is centered around a set of trajectory libraries constructed from training data. To make a forward prediction from a given time series, one interpolates the best match between other similar trajectories, and makes the corresponding interpolation between their future path.

In the extreme event case, the CCM technique is caught in a double bind. On the one hand, causal statements are *too strong* for a merely predictive context. On the other, the library look-up procedure for making forward predictions is either a

too-simplistic solution to the prediction problem, or simply the algorithmic black box begging to be replaced by any other preferred technique.

## 2.6 Mixed Methods

### 2.6.1 Overview

In some models, we have a full and explicit set of equations that model intermittent or extreme phenomena. In these cases, we can leverage these equations *in addition to* sample trajectories to understand how extreme events form, and how to predict them. Further, we can always use the equations to simulate new trajectories.

This section attempts to summarize, in particular, the work of Cousins and Sapsis (2014) [67], Farazmand and Sapsis (2017 [20], and Wan and Sapsis (2018) [64].

The basic procedure for mixed methods can be summarized as follows:

1. use the equations to determine a set of trajectories that maximizes some growth condition
2. use a pool of sample trajectories to determine the probability density of the previously identified trajectories
3. use problem-domain knowledge to identify a metric that well-classifies trajectories based on observable data

It is the first step that uses the full equations and distinguishes mixed methods from fully data driven methods. In the remainder of this section, two approaches to this first step are summarized.

### 2.6.2 Ansatz Solution

Cousins and Sapsis [67] examine the Majda-Machlaughlin-Tabak (MMT) model, a one dimensional nonlinear model of dispersive waves that exhibits a Benjamin-Feir type instability [2]. They give the MMT model as

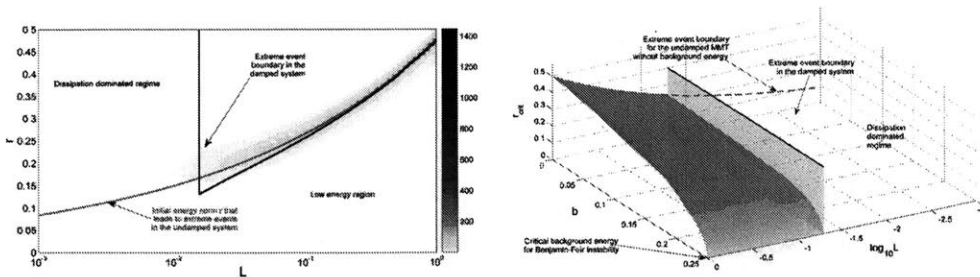


Figure 2-3: a) Plot of critical energy density  $r_{\text{crit}}(L)$  in the MMT model, as a function of length scale  $L$ . b) Plot of  $r_{\text{crit}}(L, b)$  as a function of both length scale  $L$  and background energy density  $b$ . Figures taken from Cousins and Sapsis (2014) [67], Fig. 4 and 6.

$$iu_t = |\partial_x|^\alpha u + \lambda |\partial_x|^{-\frac{\beta}{4}} (|\partial_x|^{-\frac{\beta}{4}} u)^2 |\partial_x|^{-\frac{\beta}{4}} u + iDu. \quad (2.25)$$

In the zero damping limit, they identify a family of scale invariant solutions given by

$$w_L(x, t) = \frac{1}{L^{\frac{\alpha+\beta}{2}}} u\left(\frac{x}{L}, \frac{t}{L^\alpha}\right). \quad (2.26)$$

Using this family of solutions, along with a well studied four wave interaction [16], they identify a critical local energy density  $r_{\text{crit}}(L)$ , with the same scale invariant properties. Solutions that exceed this local energy density will initiate a blow-up in finite time.

In the selectively damped case, the shape of  $r_{\text{crit}}(L)$  changes because at small length scales perturbations are dissipated before they can grow, as shown in figure 2-3.

In general, perturbations are not isolated. Therefore, the critical energy density  $r_{\text{crit}}(L)$  may *also* depend on the background energy  $b$ . The authors perform both linearized stability analysis and numerical experiments to determine the critical energy as a function of both length scale and background energy.

Together, these results can be summarized in the following: a solution with local energy density exceeding  $r_{\text{crit}}(L, b)$  will (very likely) exhibit unstable growth, and this critical energy density  $r_{\text{crit}}(L, b)$  can be determined solely by reference to 2.25.

### 2.6.3 Variational Approach

Farazmand and Sapsis (2017) [20] develop an alternative framework for using equations to predict extreme events. This variational approach seeks out an unstable manifold associated with extreme event progenitors, and then predicts extreme events based on trajectories entering this manifold.

In the general case, an intermittent set of nonlinear partial differential equations can be expressed as the following:

$$\begin{aligned} \partial_t u &= \mathcal{N}(u) \\ \mathcal{K}(u) &= 0 \\ u(\cdot, t_0) &= u_0(\cdot), \end{aligned} \tag{2.27}$$

where  $\mathcal{N}$  and  $\mathcal{K}$  are some (nonlinear) differential operators that represent time evolution and boundary conditions, respectively.

The variational approach is characterized by the following steps:

- define some scalar observable  $I(u(t))$
- look for trajectories  $u_0(t)$  that maximizes this observable
- subject to feasibility constraints

More formally, we can express this program as the solution to the following problem:

$$\begin{aligned} u_0 &= \operatorname{argsup}_{u_0 \in X} [I(u(t_0 + \tau)) - I(u(t_0))] \\ u_0(t) &\text{ satisfies equations 2.27} \\ C(u_0) &= c_0. \end{aligned} \tag{2.28}$$

The first line in equation 2.28 requires that the trajectory correspond to extreme events, or at least the sort of locally maximal events commonly associated with intermittent behavior. The second line requires that the trajectories be physical, and the third requires that they be ‘realizable,’ in problem-specific sense that corresponds to the requirement that likelihood of the given trajectory be nonvanishing.

The authors apply this approach, along with a functional optimization toolset, to the Kolmogorov Flow model of 2D incompressible fluid flow.

### 2.6.4 Closure Models

A different equation based approach is exhibited by Wan and Sapsis (2018) [64] in the context of particle-in-fluid motion. Here, the authors construct a blended slow manifold model to approximating the Maxey-Riley equation of motion [39]. This model combines a leading order perturbative approximation to the velocity term along with a data-driven closure term:

$$\begin{aligned}
 \dot{\mathbf{x}}(t) &= \mathbf{v}^*(t) + \mathbf{v}_d(t) \\
 \mathbf{v}^*(t) &= \mathbf{u}(\mathbf{x}, t) + \epsilon \left( \frac{3}{2}R - 1 \right) \frac{D\mathbf{u}(\mathbf{x}, t)}{Dt} \\
 \mathbf{v}_d(t) &= G(\xi(t), \xi(t - \tau), \xi(t - 2\tau), \dots).
 \end{aligned}
 \tag{2.29}$$

The leading order perturbative approximation  $\mathbf{v}^*(t)$  will be inaccurate when either the parameter  $\epsilon$  grows or when the flow field advective derivative  $\frac{D\mathbf{u}(\mathbf{x}, t)}{Dt}$  is large. In both cases, machine learning techniques—here, a Long Short-Term Memory (LSTM) Recurrent Neural Network (RNN)—are used to learn the data-model mismatch term  $\mathbf{v}_d(t)$ .

### 2.6.5 Point of Departure

The three approaches described above are all linked by requires a set of equations to describe the system: either a complete set, or at least a leading order approximation.



Some intermittent systems of interest, however, do not have equations that adequately describe them. This may be either because the best approximate equations simply do not reproduce extreme events, such as options pricing and the Black-Scholes model [10], or because the underlying system is poorly understood, such as in biological systems like cell adaptation [5].

## 2.7 Optimization Problems

### 2.7.1 Overview

Optimization Theory is a well-studied branch of mathematics, commonly covered in both undergraduate and graduate curricula. In this section, we will only briefly recapitulate the problem formulation, and then examine two approaches.

In an optimization problem, we seek a vector  $\mathbf{x} \in \Omega$  that extremizes (without loss of generality, minimizes) the scalar function  $f(\mathbf{x})$ . That is, we seek

$$\mathbf{x}_0 = \operatorname{argmin}_{\mathbf{x} \in \Omega} f(\mathbf{x}). \quad (2.30)$$

For simple problems, such as the linear function  $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$ , this can be done by directly solving

$$\nabla_{\mathbf{x}} f(\mathbf{x}) = 0, \quad (2.31)$$

and exhaustively searching through the solutions  $\mathbf{x}_k$ . If  $\Omega$  is closed, this direct search may further involve a recursive search through the boundary space  $\partial\Omega$ .

In the more general case, however, closed form solutions to 2.31 do not exist. In those cases, some iterative algorithm is used to select a sequence of trial points  $\{\mathbf{x}_k\}$  that hopefully converges to the global minimum.

The simplest case is when the function  $f(\mathbf{x})$  is convex. In this case, any local minimum is also a global minimum, which greatly facilitates the search. In the more general, non-convex case, some additional work must be done to ensure that the local minimum found is also a global minimum.

There is a wealth of optimization algorithms in the literature, and the choice between them is problem dependent. Some of the factors relevant to the decision include:

- Dimension of  $\Omega$
- ‘Size’ and ‘shape’ of  $\Omega$
- Linearity or non-linearity of  $f(\mathbf{x})$
- Smoothness of  $f(\mathbf{x})$
- Convexity of  $f(\mathbf{x})$
- Ease of calculating  $f(\mathbf{x})$
- Deterministic or stochastic nature of  $f(\mathbf{x})$
- Existence of analytic gradient  $\nabla_{\mathbf{x}}f(\mathbf{x})$
- Existence of analytic Hessian  $\mathbf{H}_f(\mathbf{x})$

## 2.7.2 Gradient Methods

The most common family of iterative algorithms uses the local gradient  $\nabla_{\mathbf{x}}f(\mathbf{x}_k)$  to determine the trial point  $\mathbf{x}_{k+1}$ .

The simplest example from this family is the method of steepest descent. In this method, the iterates are given by

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \gamma_k \nabla f(\mathbf{k}), \quad (2.32)$$

where  $\gamma_k$  is a step size that may depend on the step count  $k$ , the local value of the function  $f(\mathbf{x}_k)$  (or its gradient  $\nabla_{\mathbf{x}}f(\mathbf{x})$ , or Hessian  $\mathbf{H}_f(\mathbf{x})$ ), or something else.

Steepest descent is a slow first order method, and it has well-studied issues such as a tendency to zig-zag near certain kinds of extrema. Various extensions such as momentum terms have been proposed. Additionally, convergence can be improved for

many kinds of problems: for instance, smooth problems with higher order derivatives (Hessian) can be exploited to converge faster.

The method of steepest descent is guaranteed to converge only to a *local minimum*. One way to attempt to achieve global minimization is to repeat the algorithm from many initial guesses  $\mathbf{x}_0$ . Each local minima has a different basin of attraction; by judicious sampling of enough starting points it is *likely* one sequence will converge to the global minimum.

One well-known extension to the gradient descent methods, called Stochastic Gradient Descent (SGD) makes use of a stochastic approximation  $\mathcal{E}[g(\mathbf{x})] = \nabla_{\mathbf{x}}f(\mathbf{x})$ . This procedure arises frequently in the context of machine learning big data, where  $f$  is really a function of some distribution  $\mathcal{Y}$  but can only be approximated by random samples  $\{y\} \in \mathcal{Y}$ , or where  $f(\mathbf{x}) = \sum_{j < J} f_j(\mathbf{x})$  for some very large  $J$ .

The major limitation of gradient-based methods is that they require a ‘smooth enough’ function with an ‘easy to calculate’ gradient. For a low dimensional space  $\Omega$ , the gradient can be approximated by a finite difference method. As the dimension of  $\Omega$  grows, however, this quickly becomes prohibitive.

### 2.7.3 Interpolation Search

Another way to optimize  $f(\mathbf{x})$  without a gradient is to replace  $f(\mathbf{x})$  with a *surrogate*  $g(\mathbf{x})$  with the properties that:

- $g(\mathbf{x}_k) \approx f(\mathbf{x}_k)$  for every point that  $f(\mathbf{x}_k)$  is known
- $g(\mathbf{x})$ ,  $\nabla g(\mathbf{k})$ , etc. are ‘easy’ to evaluate [6][63]

In surrogate optimization,  $g(\mathbf{x})$  can be thought of as an interpolation of the true response function  $f(\mathbf{x})$ . There are many approaches to this interpolation, included Kriging models and support vector regression models. The Matlab software implementation uses a global interpolation uses a radial basis function approximation [65].

What all surrogate methods have in common is that they query the surrogate function  $g(\mathbf{x})$  in order to minimize the number of function evaluations of  $f(\mathbf{x})$ . Be-

cause this interpolation step adds significant overhead, surrogate search methods are most appropriate when evaluations of  $f(\mathbf{x})$  are expensive.

Surrogate optimization owes its origins to sensitivity analysis and optimal experiment design. Because of this, the algorithm can be seen to balance two competing goals: exploration and exploitation. Exploration is the desire to globally sample the space  $\Omega$  to find promising regions. Emphasis on exploration helps to avoid the search getting trapped in local optima, and is usually formulated in terms of maximally reducing the global uncertainty of the surrogate  $g(\mathbf{x})$ . Exploitation is the desire to search  $\Omega$  near the current best guess for the local optimum. Emphasis on exploitation is usually formulated as choosing points near the global minimum of  $g(\mathbf{x})$ , which is the ‘best guess’ for the global minimum of  $f(\mathbf{x})$ .

Surrogate search algorithms are typically divided into two stages. In the first stage, points are pseudo-randomly sampled across the entire space  $\Omega$  without regard to the shape of  $f(\mathbf{x})$ . A common strategies in this stage, especially when  $\Omega$  is high dimensional, is to sample points according to a Latin hypercube.

In the second adaptive stage, the competing goals of exploration and exploitation are balanced. In schematic form, the procedure can be expressed as:

- Use the surrogate  $g_{k-1}(\mathbf{x})$  to pick a trial point  $\mathbf{x}_k$
- Use the function  $f(\mathbf{x})$  to evaluate  $f(\mathbf{x}_k)$
- Use the pair  $(\mathbf{x}_k, f(\mathbf{x}_k))$  to update the surrogate  $g_k(\mathbf{x})$

To help avoid spurious local minima, surrogate search may employ a ‘surrogate reset’ step. After the adaptive phase has converged, all second stage points are discarded, and an additional set of new pseudo-random points are sampled. Afterwards, a new second stage is begun.

While the surrogate function  $g(\mathbf{x})$  is deterministic, there has been some work in extending this approach to stochastic functions  $f(\mathbf{x}; \omega)$ , see for instance, Sankaran et al. (2010) [49].

## 2.8 Postmortem: Discarded Attempt

### 2.8.1 Overview

This section contains a description and post-mortem reflection of a previous attempt to address the problem of extreme event prediction. Sections 2.8.2, 2.8.3, and 2.8.4 present the absolute discrimination in a context-free setting. Section 2.8.5 examines the challenges facing this approach, and explains why it turned out to be a dead-end.

### 2.8.2 Retrospective Stencil

For a more full treatment of this issue, see section 4.3.

Evaluating the goodness a predictor  $B$  requires the joint pdf  $p(A, B)$  of paired predictor-indicator values, and with finite data this joint pdf is approximated as a histogram of  $(A_k, B_k)$  pairs.

A large value of  $B$  at a certain point  $(\mathbf{x}_0, t_0)$  normally has a certain amount of fuzziness associated with it: we would call the prediction good if there were a true extreme event *near*  $(\mathbf{x}_0, t_0 + \tau)$ . Further, the paired  $(A_k, B_k)$  data may be different if we take predictions at each point and look for corresponding extreme events, or if we take events at each point and look for corresponding predictions.

In many applications, we value recall more highly than precision. Therefore, to construct the retrospective histogram of  $(A_k, B_k)$ , for every point in the domain we take the local indicator value,  $A_k$ , and pair it with the highest value of the predictor within a retrospective region,  $B_k$ . The retrospective region is the cylinder with height  $\Delta t$  and radius  $\Delta x$  centered on the point  $(\mathbf{x}_k, t_k - dt)$ .

### 2.8.3 Absolute Discrimination

The histogram of  $(A_k, B_k)$  is a discrete approximation to the joint probability distribution  $p(A, B)$ . This joint pdf has both a marginal distribution  $p(B)$  and a conditional distribution  $p(B|A)$ . The absolute difference between the mean  $\mu_B = \int Bp(B)dB$

and the conditional mean  $\mu_{B|A} = \int Bp(B|A)dB$  is the **local absolute retrospective discrimination**, and is given by

$$d_r(A) = |\mu_{B|A} - \mu_B| = \left| \int Bp(B|A)dB - \int Bp(B)dB \right|. \quad (2.33)$$

The local absolute retrospective discrimination measures the difference in the mean of the distribution of best predictor values  $B$  given knowledge of the associated indicator  $A$ . This retrospective viewpoint is best suited to emphasize the pairs  $(A_k, B_k)$  that actually correspond to extreme events—and to see if they have extreme predictions.

The **absolute retrospective discrimination**, defined as

$$D_r = \sum_A d_r(A) = \int |\mu_{B|A} - \mu_B|dA, \quad (2.34)$$

is the area between the curve  $f(A) = \mu_{B|A}$ , the average predictor for events of extreme value  $A$ , and  $\mu_B$ , the mean predictor for all events.

$D_r$  is maximized when  $\mu_{B|A} \neq \mu_B$  for ‘many’ values of  $A$ . That is to say, for ‘many’ values of  $A$ , knowledge of  $A$  gives additional knowledge about the distribution of  $B$ . This is similar to the statement that relative entropy from  $p(B)$  to  $p(B|A)$  is positive, though  $D_r$  only tracks the mean and not the higher moments of the distribution.

In practice, the overwhelming number of not-very-extreme-events means that  $\mu_{B|A} \approx \mu_B$  except when  $A$  takes extreme values. If  $B$  is a good predictor, then  $B$  will also take extreme values,.

#### 2.8.4 Interpretation

The quantity  $D_r$  does not correspond directly to the recall of  $B$  for  $A$ , and its absolute magnitude does not have any particular physical interpretation. Rather, for two potential predictors  $B^{(1)}$  and  $B^{(2)}$ , the relative magnitudes of  $D_r^{(1)}$  and  $D_r^{(2)}$  correspond to the relative goodness of each predictor for extreme values of  $A$ . In the examples below, the predictor from a specified class that maximizes  $D_r$  will be

interpreted as the best predictor.

### 2.8.5 Challenges

The absolute discrimination faces three major types of challenges. First, because it can be understood as a type non-parametric regression, it faces all of the associated challenges. In particular, relatively small extreme event counts leads to imprecise estimation of  $\mu_{B|A}$  for extreme  $A$ .

Second, the explicit computation of quantities like  $\mu_B$  cause the precise measure properties of the training data to become important. As discussed later in section 4.3, this is a complicated issue, and a robust metric ought be less sensitive to the technical choices in this area.

Third,  $D_r$  does not have an easy interpretation as anything other than a raw ranking of different predictor choices. This requires a fallback to other quantities, like total accuracy, in order to characterize a predictors quality in an absolute sense. Additionally,  $D_r$  and the entire  $\mu_{B|A}$  graph fail to offer nuanced information about the false positive false negative trade-off decision, instead hiding the issue in the choice of retrospective stencil.

Finally, equations 2.33 and 2.34 are not invariant under rescalings of either the true indicator function  $A$  or the predictor function  $B$ . While linear transformations may be swallowed by normalization conditions, the different choices of amplitude or energy ( $|A|$  or  $|A|^2$ ) will lead, in general, to different optimal predictors.

# Chapter 3

## Receiver Operating Characteristic Metrics

### 3.1 Problem Overview

In order to overcome the problems inherent in applying standard loss function formulations to extreme events, in this report we will develop an alternate methodology. This methodology, though based on binary classification into quiescent and extreme bins, will be based on an extension to the concept of receiver operating characteristic curves. As a result, this method will not depend on particular choice of threshold, either of the indicator or predictor.

We will start by identifying a set of derived quantities, precision and recall, that are better suited than the total error rate to the case of extreme events. Next, we will show how to use these quantities to derive a receiver operating characteristic curve whose integral has a straightforward interpretation as a performance metric that is independent of the choice of predictor threshold.

We will then show how to straightforwardly generalize this curve into a surface containing all possible choices of threshold for both indicator and predictor. We will point out features of this surface, including a knuckle that corresponds to scale separation in the underlying system. Finally, we will use features of this surface, including its integral, to construct a collection of metrics that are well suited to



evaluating predictors.

The last section of this report will demonstrate the application of this methodology to three test scenarios, including a system with scale separation, and one that is fundamentally linearly inseparable.

## 3.2 Groundwork

### 3.2.1 Basic Quantities

A training data set takes the form of a set of pairs  $S = \{(a_i, b_i)\}$ , where  $a_i$  is the indicator of the  $i^{\text{th}}$  data point, and  $b_i$  is the corresponding predictor. The set of these predictor-indicators pairs together defines a two dimensional joint distribution, with probability density function (pdf)  $f_{ab}(a, b)$  and cumulative integral function

$$F_{ab}(\hat{a}, \hat{b}) = P(a_i > \hat{a}, b_i > \hat{b}). \quad (3.1)$$

The function  $F_{ab}(\hat{a}, \hat{b})$  may be interpreted as the probability that a give pair lies above both a threshold  $\hat{a}$  along the indicator dimension and  $\hat{b}$  along the predictor dimension. We use hat notation when the choice of  $\hat{a}$  or  $\hat{b}$  corresponds implicitly to a threshold in this sense. A value exceeding the threshold is called extreme, and a value not exceeding it is called quiescent.

The function  $F_{ab}(\hat{a}, \hat{b})$  is similar to, but not the same as the cumulative distribution function (cdf)  $P(a_i < \hat{a}, b_i < \hat{b})$ . The given form of integrated probability is chosen for this work, because the prediction problem privileges the ‘quadrant’ of the pdf containing true positives, and derivatives of this integral will appear.

The cumulative integral function can be related to the cdf and its marginals by the formula

$$P(a_i > \hat{a}, b_i > \hat{b}) = 1 - P(a_i < \hat{a}) - P(b_i < \hat{b}) + P(a_i < \hat{a}, b_i < \hat{b}) \quad (3.2)$$

Henceforth, we will only refer to the cumulative integral function  $F_{ab}(\hat{a}, \hat{b})$ .

The joint distribution may be constructed from simulations, from empirical measurements, or from analytical models. We will use the term **histogram** in this report to refer to the data and its functional representation as a probability distribution.

A fixed choice of  $\hat{a}$  and  $\hat{b}$  defines a binary classification with four possibilities:

- True Positive – an event predicted to be extreme that is actually extreme
- True Negative – an event predicted to be quiescent that is actually quiescent
- False Positive – an event predicted to be extreme that is actually quiescent
- False Negative – an event predicted to be quiescent that is actually extreme

From this classification, we will define the following three important quantities:

**Definition 3.2.1** (Derived Quantities). The **precision**, denoted by  $s$ , is the probability that an event is a true positive, given that it is predicted to be extreme. The **recall** (sometimes called specificity), denoted by  $r$ , is the probability that an event is a true positive, given that it is actually extreme. The **extreme event rate**, denoted by  $q$ , is the probability that an event is actually extreme.

In probabilistic notation, these quantities may be expressed as:

$$\begin{aligned}
 s(\hat{a}, \hat{b}) &= P(a > \hat{a} | b > \hat{b}) \\
 r(\hat{a}, \hat{b}) &= P(b > \hat{b} | a > \hat{a}) \\
 q(\hat{a}) &= P(a > \hat{a}).
 \end{aligned} \tag{3.3}$$

The precision, recall, and extreme event rate may be expressed in terms of the cumulative integral function (and its marginals) using the following helper functions:

$$\begin{aligned}
 D(\hat{a}, \hat{b}; F_{ab}) &= F_{ab}(\hat{a}, \hat{b}) \\
 E(\hat{b}; F_{ab}) &= F_{ab}(-\infty, \hat{b}) = F_b(\hat{b}) \\
 F(\hat{a}; F_{ab}) &= F_{ab}(\hat{a}, -\infty) = F_a(\hat{a}).
 \end{aligned} \tag{3.4}$$

Using these helper functions, the derived quantities may be written as

$$\begin{aligned}
s(\hat{a}, \hat{b}; F_{ab}) &= \frac{D(\hat{a}, \hat{b}; F_{ab})}{E(\hat{b}; F_b)} \\
r(\hat{a}, \hat{b}; F_{ab}) &= \frac{D(\hat{a}, \hat{b}; F_{ab})}{F(\hat{a}; F_a)} \\
q(\hat{a}; F_{ab}) &= F(\hat{a}; F_a).
\end{aligned} \tag{3.5}$$

From hereon, the functional dependence of quantities on the cumulative integral function  $F_{ab}$  will be suppressed except when relevant.

### 3.2.2 Basic Properties

By construction, the derived quantities have the following monotonicity properties:

**Theorem 2** (Monotonicity). *The precision,  $s(\hat{a}, \hat{b})$ , is monotonic in its first argument,  $\hat{a}$ . The recall,  $r(\hat{a}, \hat{b})$ , is monotonic in its second argument,  $\hat{b}$ . The ferocity,  $q(\hat{a})$ , is monotonic in its first argument,  $\hat{a}$ .*

Additionally, the choice of precision and recall as binary classification metrics is motivated by the following three invariances:

**Theorem 3** (Invariance-I). *Precision, recall, and extreme event rate depend only on the shape of  $F_{a,b}$*

**Theorem 4** (Invariance-II). *Let  $f_1(a, b)$  and  $f_2(a, b)$  be two histograms with the property that, outside of a certain region  $[-\infty, a^*] \otimes [-\infty, b^*]$ ,  $f_1(a, b) = \beta f_2(a, b)$  for some fixed constant  $\beta$ . That is to say,  $f_1(a, b)$  and  $f_2(a, b)$  correspond to histograms that differ only by some number of points  $(a_i, b_i)$  for which  $a_i < a^*$  and  $b_i < b^*$ .*

*Then for all  $\hat{a} > a^*$  and  $\hat{b} > b^*$ ,  $s(\hat{a}, \hat{b}; F_1) = s(\hat{a}, \hat{b}; F_2)$  and  $r(\hat{a}, \hat{b}; F_1) = r(\hat{a}, \hat{b}; F_2)$ .*

**Theorem 5** (Invariance-III). *Let  $h_1(y), h_2(y) : \mathcal{R} \rightarrow \mathcal{R}$  be order-preserving monotonic functions. Let  $F_{ab}$  be a histogram, and let  $F_{xy}$  be the histogram formed from the data set  $\{(h_1(a_i), h_2(b_i))\}$ . Then*

$$\begin{aligned}
s(\hat{a}, \hat{b}; F_{ab}) &= s(h_1(\hat{a}), h_2(\hat{b}); F_{xy}) \\
r(\hat{a}, \hat{b}; F_{ab}) &= r(h_1(\hat{a}), h_2(\hat{b}); F_{xy}) \\
q(\hat{a}; F_{ab}) &= q(h_1(\hat{a}); F_{xy}).
\end{aligned} \tag{3.6}$$

*In other words, precision, recall, and extreme event rate are invariant under arbitrary nonlinear rescalings of the indicator and predictor.*

*Proof.* The proof of 3 follows directly from the definitions of  $q$ ,  $r$ , and  $s$  in 3.3, where they are expressed in terms of conditional probabilities.

The proof of 4 follows from the restatement of definition 3.3 in the form of the helper functions in equations 3.4 and 3.5. The cif  $F_{ab}$  may be expressed as a definite integral of the pdf  $f_{ab}$ , with limits as given in the statement of the theorem. Therefore, any differences in the histogram outside of the limits of integration cannot impact the value of the definite integral.

Because  $s$  and  $r$  are ratios of definite integrals, each of which is linear in its integrand, they are invariant under the linear rescaling  $\beta$ . Note that this is *not* true for  $q$ , which is not defined as a ratio.

Finally, the proof of 5 follows from the familiar u-substitution rule of ordinary calculus, restricted to a class of particularly well behaved substitutions.  $\square$

The invariance in theorem 3 justifies the ambivalence we will display between pdf formulation and histogram formulations in the discussion to come.

The invariance of theorem 4 is directly aimed at the preference for extreme events. A predictor designed to predict extreme events should not be sensitive to very quiescent training data events. This invariance may justify pruning the data, such as a resampling technique that preferentially discards quiescent events.

Note that while precision and recall are invariant, the extreme event rate is not. That is to say, by removing quiescent events from the data set, the apparent fraction of extreme events will increase.

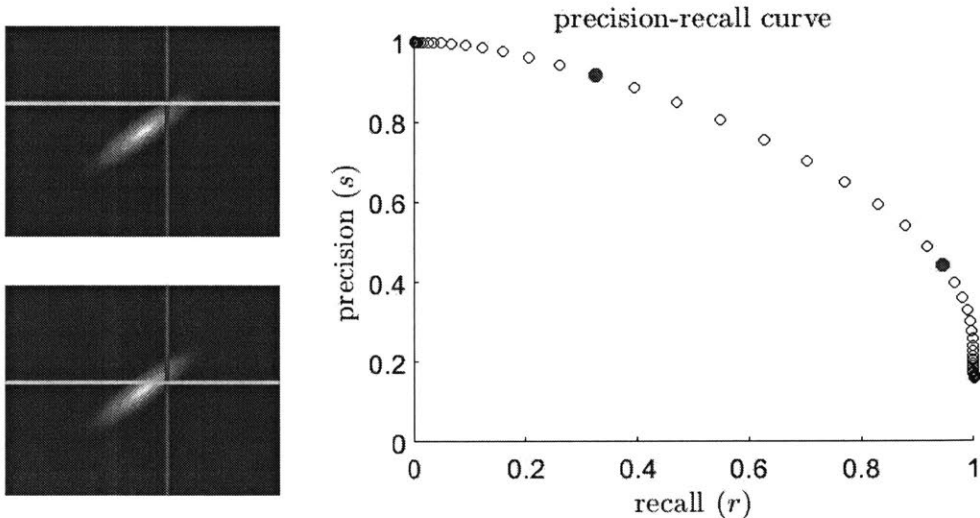


Figure 3-1: Plot of a) a sample pdf and b) the corresponding precision recall (SR) curve. Each circle corresponds to a  $(\hat{a}, \hat{b})$  pair: the curve is generated by fixing  $\hat{a}$  and letting  $\hat{b}$  vary.

The final invariance (5), minimizes the issue of scale in the choice of the indicator and predictor functions. For instance, an indicator for extreme waves may be wave height ( $\approx x$ ) or energy ( $\approx x^2$ ). However, because of the monotonic relationship between height and energy, the choice won't effect derived quantities further down the line. This is particularly useful in limiting the hypothesis space of potential predictors.

### 3.3 QRS Surface

#### 3.3.1 Precision Recall Curve

Precision, recall, and extreme event rate are dependent on the choice of  $\hat{a}$  and  $\hat{b}$ , together the choice of thresholds define a set of possible s-q-r tuples.

**Definition 3.3.1** (Precision Recall Curve). For a fixed  $\hat{a}$ , the **precision recall curve** (SR curve) is the parametric curve defined by

$$\rho(\hat{b}; \hat{a}) = (r(\hat{a}, \hat{b}), s(\hat{a}, \hat{b})). \quad (3.7)$$

Because  $r(\hat{a}, \hat{b})$  is invertible in its second argument (theorem 2), this curve can also be interpreted to express precision as a unique function of recall and the extreme event rate:

$$s = s(r; q). \tag{3.8}$$

An example SR curve (for fixed  $q$ ) is exhibited in figure 3-1. A few features stand out. First, smaller values of recall correspond to larger values of precision, and vice versa. This is intuitive: in order to be sure to catch every extreme event (high recall), the predictor will have to let through many false positive quiescent events (low precision).

Second, this SR curve is monotonic. This feature is not guaranteed by construction. However, it is a typical feature because for a good predictor  $P(a > \hat{a}|b)$  is ‘mostly’ monotonic increasing in  $b$ . Any SR curve  $\rho_1$  that is not monotonic can be transformed into a new SR curve  $\rho_2$  such  $\rho_2$  is monotonic and  $\rho_2$  dominates  $\rho_1$  (see appendix 3.7).

Third, the precision doesn’t fall to 0 with increasing  $r$ , even when the predictor threshold ( $\hat{b}$ ) is arbitrarily small. Instead, the following limit obtains:

**Theorem 6** (Extreme Event Rate Correspondence). *Let  $F_{ab}$  be a histogram with SR curve  $\rho$  corresponding to extreme event rate  $q$ . Then*

$$\lim_{r \rightarrow 1} s(r; q) = q. \tag{3.9}$$

### 3.3.2 Precision Recall Curve Metric

The SR curve can be used to measure the quality of a predictor independently of the choice of predictor threshold  $\hat{b}$ . The ideal predictor’s SR curve would run from  $(0, 1)$  to  $(1, 1)$ , and then down from  $(1, 1)$  to  $(q, 1)$ .

One way measure the predictor quality would be to take the distance of closest approach between the curve and the point  $(1, 1)$ . This method, which sees some use in evaluating precision recall curves, is difficult to generalize to the 3D context of

variable extreme event rate.

Another way to evaluate the SR curve is given below:

**Definition 3.3.2** (Area Under the Curve). The **Area Under the Curve** ( $\alpha$ ) is the area under the SR curve corresponding to indicator threshold  $\hat{a}$  (alternatively  $q(\hat{a})$ ), and may be expressed as

$$\begin{aligned}\alpha(\hat{a}) &= \int_0^1 s(r) dr \\ &= \int_{-\infty}^{\infty} s(\hat{b}) \left| \frac{\partial r}{\partial \hat{b}} \right| d\hat{b}.\end{aligned}\tag{3.10}$$

For two predictor histograms  $f_1$  and  $f_2$ , with corresponding SR curves  $\rho_1$  and  $\rho_2$ , if  $\forall r \in [0, 1] s_1(r) \geq s_2(r)$ , then  $\alpha_1 \geq \alpha_2$ , with strict inequality if  $s_1(r) > s_2(r)$  over some interval of finite measure. However, the converse is *not* true: a SR curve  $\rho_1$  may exceed  $\rho_2$  over some intervals but not others, and vice versa.

### 3.3.3 Precision Recall Rate Surface

The area under the curve  $\alpha$  evaluates a predictor without making an explicit choice of  $\hat{b}$ . We can take this process one step further to remove dependence on  $\hat{a}$ .

**Definition 3.3.3** (QRS Surface). The **precision recall rate surface** (QRS surface) is the parametric surface defined by

$$\sigma(\hat{a}, \hat{b}) = (r(\hat{a}, \hat{b}), s(\hat{a}, \hat{b}), q(\hat{a})).\tag{3.11}$$

Like the SR curve,  $q$  and  $r$  may be inverted sequentially to expressed the surface as a function given by

$$s = s(r(\hat{a}, \hat{b}), q(\hat{a})).\tag{3.12}$$

qrs plot: bimodal ( $\gamma = 0.05, \rho = 2.0$ )

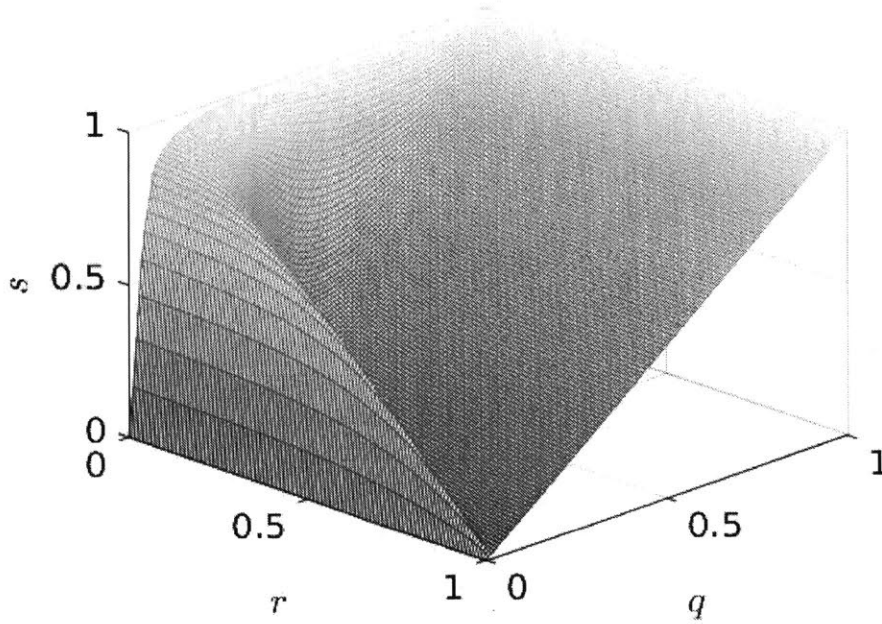


Figure 3-2: Plot of a sample Precision Recall Extreme Event Rate (QRS) Surface. The surface is generated by varying  $\hat{a}$  and  $\hat{b}$ .

### 3.3.4 Precision Recall Rate Surface Metric

By analogy with the  $\alpha$  functional on the SR curve, we can define an enclosed volume functional for the QRS surface as well. And, like the  $\alpha$ , this volume under the surface also partially orders predictors by quality.

**Definition 3.3.4** (Volume Under the Surface). The **Volume Under the Surface** ( $V$ ) is the volume under the precision recall rate surface, and may be expressed as

$$V = \int_0^1 \int_0^1 s(r, q) dr dq \tag{3.13}$$

Equation 3.13 can be rewritten in terms of thresholds and the helper functions from equation 3.4 as

$$V = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{D}{E} \Big| \frac{1}{F} \frac{\partial D}{\partial \hat{b}} \frac{\partial F}{\partial \hat{a}} \Big| d\hat{a} d\hat{b}. \tag{3.14}$$



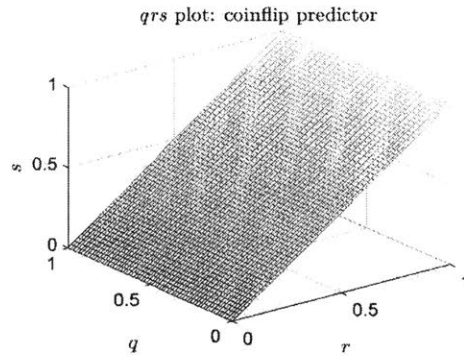


Figure 3-3: Plot of the QRS Surface for the coinflip predictor.  $V = 0.5$ .

## 3.4 QRS Features

### 3.4.1 Coinflip Predictor

The coinflip predictor is the naive predictor that is completely independent of the indicator. Its characteristic triangular shape, shown in 3-3, has the minimum value of the Volume Under the Surface:  $V = 0.5$ .

Any predictor for which  $V < 0.5$  must contain some nonzero mutual information between predictor and indicator, but is such that a simple thresholding classification fails. Further, when the QRS surface for some predictor dips below the coinflip, it implies that the relationship between indicator and predictor is not monotonic.

The most simple explanation for such failure is that the predictor is inverted: that is, the binary classification assumes that large  $B$  predict extreme events, while in fact small  $B$  better predict extreme events. Section 3.6.3 exhibits one such histogram, and explains how a better predictor might be constructed.

### 3.4.2 Knuckle

Figure 3-2 exhibits one type of feature that may be present on the QRS surface: the knuckle.

In figure 3-2, the knuckle is the ridge of low  $q$  and high  $s$ . Figure 3-4 exhibits a fixed  $r$  slice of the QRS surface—the QS curve. On this curve, the knuckle is visible as a local maximum/minimum pair.

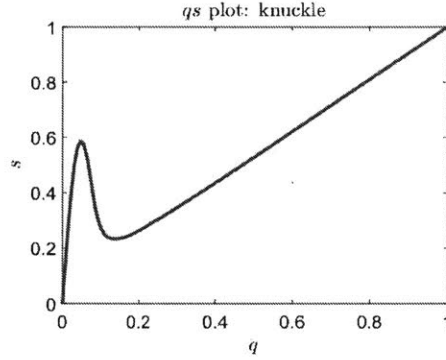


Figure 3-4: Precision-Rate slice of the QRS plot in figure 3-2, where  $r = 0.5$ . The knuckle is captured as the non-decreasing interval just past  $q = 0.1$ .

As a preliminary matter, we will recapitulate several results that limit the possible shapes of the QS curve.

1.  $0 \leq s \leq 1$
2.  $s(q = 0) = 0$
3.  $s(q = 1) = 1$
4.  $\frac{\partial s}{\partial q}(q = 0) \geq 0$
5.  $\frac{\partial s}{\partial q}(q = 1) \geq 0$

Property 1 is a simple consequence of the definition of precision as a conditional probability. Properties 2 and 3 require somewhat more machinery.

**Theorem 7** (Ferocity Limit Theorem). *Let  $F_{ab}$  be a histogram such that  $f_{ab}$  is continuous and has finite support. Then  $\forall r$ ,  $s(r, q = 0) = 0$  and  $s(r, q = 1) = 1$ .*

*Proof.* Consider the  $q = 1$  case.

Following equation 3.5,  $s = \frac{D(\hat{a}, \hat{b})}{E(\hat{b})}$  and  $q = F(\hat{a})$ . Because  $q = 1$ ,  $F(\hat{a}) = 1$ ,  $\hat{a}$  must be less than the smallest value of  $a$  associated with the support of  $f_{ab}$ .

Because  $f_{ab}$  has no support for  $a < \hat{a}$ , all integrals of the form  $\int_{\hat{a}}^{\infty}$  may be replaced by  $\int_{-\infty}^{\infty}$  without change in value.

This means we may expand  $D(\hat{a}, \hat{b})$  to rewrite  $s$  as  $\frac{E(\hat{b})}{E(\hat{b})} = 1$ .

The  $q = 0$  case proceeds similarly, with flipped inequality of support. □

Properties 4 and 5 follow directly from the continuity requirement that  $s$  reach its values in properties 2 and 3 without exceeding the bounds set by property 1.

Together, these conditions restrict the possible shapes of the QS curve  $s(q)$ . There can be no knuckles when  $\frac{\partial^2 s}{\partial q^2}$  has zero roots. The partial derivative  $\frac{\partial^2 s}{\partial q^2}$  must have at least one root, in order for  $\frac{\partial s}{\partial q}$  to become negative.

### Second Derivative: Identically Zero

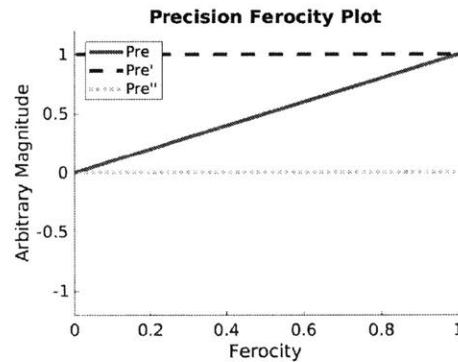


Figure 3-5: Sample PF plot when the second derivative  $\frac{\partial^2 s}{\partial r^2}$  vanishes.

The first case is when the partial derivative  $\frac{\partial^2 s}{\partial r^2}$  vanishes. Because the Precision is fixed at  $r = 0$  and  $r = 1$ , there is only one possible curve:  $s = r$ . This is simply the coinflip predictor.

### Second Derivative: Zero Roots

When the second derivative has zero roots, there are two possible shapes:

$\frac{\partial^2 s}{\partial r^2} > 0$  First,  $\frac{\partial^2 s}{\partial r^2}$  is always positive. In this case, the PF curve is always *concave upwards*. Because it is everywhere concave upwards, it lies below the coinflip line (magenta dash-dot). If this shape obtains for all values of recall, then  $VUS < 0.5$ .

Even if  $\frac{\partial^2 s}{\partial r^2} > 0$  only obtains for some values of the recall, there must still be points (probably along the always concave PF curve) at which  $s < r$ .

Together, these results imply that no good predictor may have  $\frac{\partial^2 s}{\partial r^2} > 0$  along an entire PF curve.

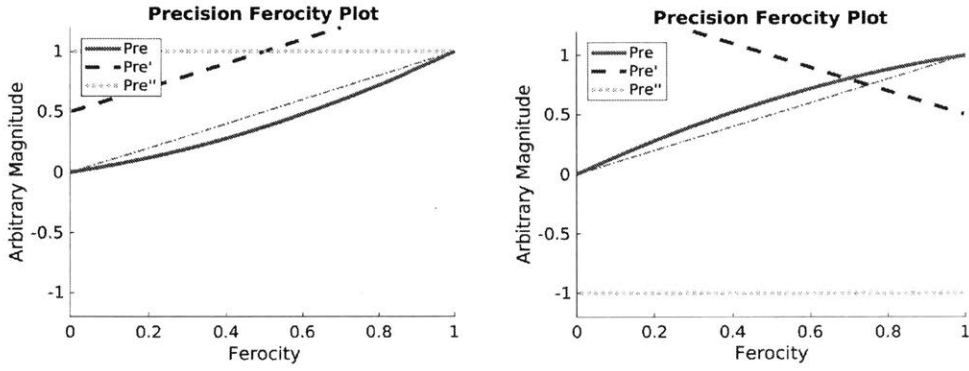


Figure 3-6: Sample PF plot when the second derivative  $\frac{\partial^2 s}{\partial r^2}$  is a) always positive and b) always negative.

$\frac{\partial^2 s}{\partial r^2} < 0$  Second,  $\frac{\partial^2 s}{\partial r^2}$  is always negative. In this case, the PF curve is concave down, and it is the case that  $s > r \forall r$ .

This is a characteristic shape for a good predictor.

### Second Derivative: One Root

When the second derivative has one root, there are three possible shapes:

$\frac{\partial^2 s}{\partial r^2}$  **crosses from positive to negative** First, the second derivative may cross from positive to negative.

$\frac{\partial^2 s}{\partial r^2}$  **crosses from negative to positive**,  $\frac{\partial^2 s}{\partial r^2} \geq 0 \forall r$  Second, the second derivative may cross from negative to positive, but the first derivative is always positive.

$\frac{\partial^2 s}{\partial r^2}$  **crosses from positive to negative**  $\exists r \mid \frac{\partial^2 s}{\partial r^2} < 0$  Finally, the second derivative may cross from negative to positive, and the first derivative has sign changes. This is that case that leads to local extrema, and a knuckle point on the PF curve.

### Second Derivative: Multiple Roots

Because one root is enough to give rise to a knuckle, we won't consider more complicated PF curves in detail. We will only mention that multiple roots may give rise to multiple knuckles centered on different values of the Ferocity.

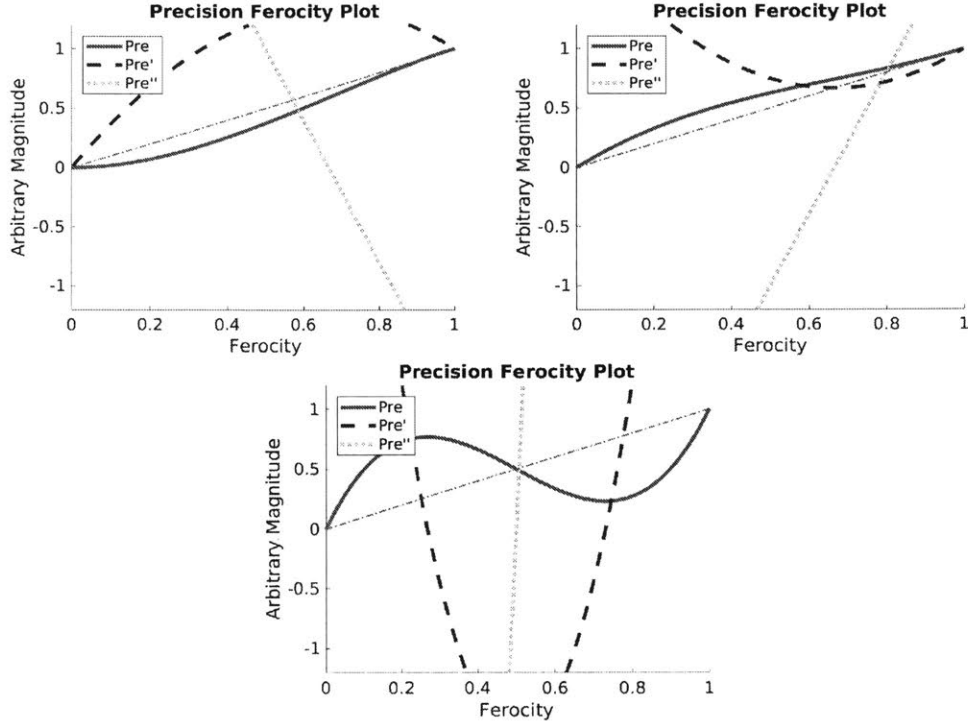


Figure 3-7: Sample PF plot when the second derivative  $\frac{\partial^2 s}{\partial r^2}$  has one root. a)  $\frac{\partial^2 s}{\partial r^2}$  is decreasing. b)  $\frac{\partial^2 s}{\partial r^2}$  is increasing, and the first derivative is always positive. c)  $\frac{\partial^2 s}{\partial r^2}$  is increasing, and the first derivative changes sign.

### 3.4.3 Histogram Correspondance

In order to express the previous derivative conditions in terms of the predictor-indicator histogram, we need to express the derivatives in terms of the histogram  $F_{ab}$  and thresholds  $\hat{a}$  and  $\hat{b}$ .

From definition 3.3.3

$$\sigma(\hat{a}, \hat{b}) = (r(\hat{a}, \hat{b}), s(\hat{a}, \hat{b}), q(\hat{a})).$$

Through parametric differentiation, the total derivatives are given by

$$\begin{aligned}
\frac{ds}{d\hat{a}} &= \frac{\partial s}{\partial r} \frac{dr}{d\hat{a}} + \frac{\partial s}{\partial q} \frac{dq}{d\hat{a}} \\
\frac{ds}{d\hat{b}} &= \frac{\partial s}{\partial r} \frac{dr}{d\hat{b}}.
\end{aligned} \tag{3.15}$$

But via definition 3.2.1 (re-expressed in equation 3.5), the total derivatives are also given by

$$\begin{aligned}
\frac{ds}{d\hat{a}} &= \frac{1}{E} \frac{\partial D}{\partial \hat{a}} \\
\frac{ds}{d\hat{b}} &= \frac{E \frac{\partial D}{\partial \hat{b}} - D \frac{\partial E}{\partial \hat{b}}}{E^2} \\
\frac{dr}{d\hat{a}} &= \frac{F \frac{\partial D}{\partial \hat{a}} - D \frac{\partial F}{\partial \hat{a}}}{F^2} \\
\frac{dr}{d\hat{b}} &= \frac{1}{F} \frac{\partial D}{\partial \hat{b}} \\
\frac{dq}{d\hat{a}} &= \frac{\partial F}{\partial \hat{a}}.
\end{aligned} \tag{3.16}$$

Putting together equations 3.15 and 3.16, we get

$$\begin{aligned}
\frac{\partial s}{\partial q} &= \frac{1}{\frac{dq}{d\hat{a}}} \left[ \frac{ds}{d\hat{a}} - \frac{\partial s}{\partial r} \frac{dr}{d\hat{a}} \right] \\
&= \frac{1}{\frac{dq}{d\hat{a}}} \left[ \frac{ds}{d\hat{a}} - \frac{\frac{ds}{d\hat{b}}}{\frac{dr}{d\hat{b}}} \frac{dr}{d\hat{a}} \right] \\
&= \frac{1}{\frac{\partial F}{\partial \hat{a}}} \left[ \frac{1}{E} \frac{\partial D}{\partial \hat{a}} - \frac{E \frac{\partial D}{\partial \hat{b}} - D \frac{\partial E}{\partial \hat{b}}}{E^2} \frac{F \frac{\partial D}{\partial \hat{a}} - D \frac{\partial F}{\partial \hat{a}}}{F^2} \right] \\
&= \frac{1}{E \frac{\partial F}{\partial \hat{a}}} \left[ \frac{\partial D}{\partial \hat{a}} - \frac{(E \frac{\partial D}{\partial \hat{b}} - D \frac{\partial E}{\partial \hat{b}})(F \frac{\partial D}{\partial \hat{a}} - D \frac{\partial F}{\partial \hat{a}})}{EF \frac{\partial D}{\partial \hat{b}}} \right] \\
&= \frac{D}{E^2 F \frac{\partial F}{\partial \hat{a}} \frac{\partial D}{\partial \hat{b}}} \left[ F \frac{\partial D}{\partial \hat{a}} \frac{\partial E}{\partial \hat{b}} + E \frac{\partial F}{\partial \hat{a}} \frac{\partial D}{\partial \hat{b}} - D \frac{\partial F}{\partial \hat{a}} \frac{\partial E}{\partial \hat{b}} \right].
\end{aligned} \tag{3.17}$$

Note that each partial is negative, but in the final line the partials always appear in pairs.

In order to create a knuckle, there must be a (a pair of) sign changes in  $\frac{\partial s}{\partial q}$ . But by equation 3.17, that partial derivative can only change sign at a root of the square bracketed quantity, and none of the ‘raw quantities’ may change sign.

The square bracketed quantity is symmetric under the exchange  $s \rightarrow r$  and  $\hat{a} \rightarrow \hat{b}$ .

Written out long form, the bracketed quantity from equation 3.17 is given by

$$\begin{aligned}
Q(\hat{a}, \hat{b}) = & \int_{\hat{a}}^{\infty} \int_{-\infty}^{\infty} f_{ab}(a, b) da db \int_{\hat{b}}^{\infty} f_{ab}(\hat{a}, b) db \int_{-\infty}^{\infty} f_{ab}(a, \hat{b}) da \\
& + \int_{-\infty}^{\infty} \int_{\hat{b}}^{\infty} f_{ab}(a, b) da db \int_{-\infty}^{\infty} f_{ab}(\hat{a}, b) db \int_{\hat{a}}^{\infty} f_{ab}(a, \hat{b}) da \\
& - \int_{\hat{a}}^{\infty} \int_{\hat{b}}^{\infty} f_{ab}(a, b) da db \int_{-\infty}^{\infty} f_{ab}(\hat{a}, b) db \int_{-\infty}^{\infty} f_{ab}(a, \hat{b}) da, \tag{3.18}
\end{aligned}$$

or in terms of probability and marginal pdfs,

$$\begin{aligned}
Q(\hat{a}, \hat{b}) = & P(a \geq \hat{a})P(b = \hat{b})P(a = \hat{a}, b \geq \hat{b}) \\
& + P(b \geq \hat{b})P(a = \hat{a})P(a \geq \hat{a}, b = \hat{b}) \\
& - P(a \geq \hat{a}, b \geq \hat{b})P(a = \hat{a})P(b = \hat{b}). \tag{3.19}
\end{aligned}$$

Typically, the prefactor  $\frac{D}{E^2 F \frac{\partial F}{\partial a} \frac{\partial D}{\partial b}}$  will be positive, and the first two terms of  $Q$  will be large and positive. That implies that  $Q$  can only change sign when the third term grows large enough in magnitude to dominate the other two terms. This growth can only occur in situations where the following **three** conditions **all** hold:

1.  $P(a = \hat{a}, b \geq \hat{b})$  is small compared to  $P(a = \hat{a})$
2.  $P(a \geq \hat{a}, b = \hat{b})$  is small compared to  $P(b = \hat{b})$
3.  $P(a \geq \hat{a}, b \geq \hat{b})$  is *not* small compared to **both**  $P(a \geq \hat{a})$  **and**  $P(b \geq \hat{b})$

These conditions are met when  $f_{ab}$  exhibits scale separation, as in figure 3-8. In the figure, the thresholds  $\hat{a}$  and  $\hat{b}$  that correspond to negative values in equation 3.17 are sketched out.

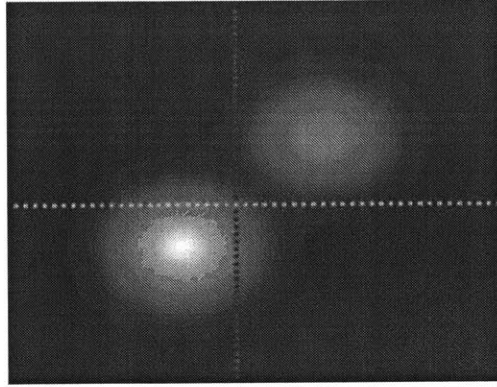


Figure 3-8: Sample pdf exhibiting scale separation. This demo is explored in section 3.6.1.

### 3.5 QRS Metrics

In this section, we will motivate a collection of metrics to assess the quality of a predictor.

The first metric is the Volume Under the Surface  $V$ , defined in equation 3.13. This quantity is large when ‘many pairs’ of  $(\hat{a}, \hat{b})$  lead to a binary classification with both high precision and high recall. It weighs contributions from different possibilities of the extreme event rate uniformly in  $q$ , and balances precision and recall in a typical manner for receiver operating characteristic curves.

However,  $V$  does not select for QRS surface knuckles. The knuckle is the characteristic feature of histograms with scale separated extreme events. We need a metric that selects for the special characteristic of a predictor that is suited for predicting extreme events.

In section 3.4.2, we defined the knuckle in terms of the partial derivative  $\frac{\partial s}{\partial q}$ . This suggests a productive avenue: the metric should select for large negative partial derivatives, which indicate large separation between the local maximum and minimum, which indicate the presence of a knuckle.

The first possibility is to selection the maximum value of  $\frac{\partial s}{\partial q}$ . In order to select for a knuckle that has large extent along  $r$ , we should add up the maxima for different values of  $r$ . Together, this gives the first knuckle metric:



$$\eta_1 = \int_0^1 \max_{q \in [0,1]} \frac{\partial s}{\partial q}(r, q) dr. \quad (3.20)$$

A second possibility is to look for the size of the gap between the maximum and minimum. The total variation, along  $q$ , is defined as

$$\eta_2 = \int_0^1 \int_0^1 \left| \frac{\partial s}{\partial q} \right| dq dr. \quad (3.21)$$

These two metrics track each other, though  $\eta_2$  is more robust numerically.

The derivative metrics have two drawbacks. First, while they select for large knuckles, they don't guarantee a good binary classification on the knuckle. That is to say, a large partial derivative  $\frac{\partial s}{\partial q}$  doesn't guarantee that the maximum  $s$  is close to 1, and a large total variation doesn't guarantee that the variation isn't due to many tiny squiggles.

Second, both metrics require calculating the partial derivative  $\frac{\partial s}{\partial q}$  explicitly. This is an involved calculation, with many potential sources of numerical error.

To overcome these difficulties, we suggest the Maximum Adjusted Area Under the Curve ( $\alpha^*$ ) as a choice of metric, given by

$$\alpha^* = \max_{q \in [0,1]} (\alpha(q) - q). \quad (3.22)$$

The quantity  $\alpha(q) - q$  is a measure, at extreme event rate  $q$ , of how much better a predictor  $B$  is than the coinflip predictor. When  $\alpha(q) - q \gg 0$ , the predictor does an excellent job of predicting extreme events at the threshold  $\hat{a}$  corresponding to the extreme event rate  $q$ . Conversely, when  $\alpha(q) - q \approx 0$  (or even  $\alpha(q) - q < 0$ ), the predictor is poor *at that extreme event rate*.

A good extreme event predictor will be a good predictor not just over possible choices of threshold (Volume Under the Surface), but at some particular thresholds that correspond to a rare rate of extreme events.

Correspondingly, the adjusted area under the curve is biased to predictors that perform well at a low extreme event rate. The precision is bounded above by  $s \leq 1$ , and the coinflip predictor has  $s = q$ . Together, that means that there is more 'room'

for a predictor to outperform the coinflip predictor at lower  $q$ . That is to say,  $\alpha^*$  selects not just for predictors that perform well as certain extreme event rates, but for predictors that perform well when extreme events are *rare*.

## 3.6 Test Scenarios

### 3.6.1 Bimodal Predictor

In order to demonstrate the effects of bimodal data on the derived quantities, we will construct a bimodal pdf, given by:

$$f_{ab}(a, b; \gamma, \rho) = \frac{1}{\beta} [\exp(-(a^2 + b^2)\rho^2) + \gamma \exp(-((a - 1)^2 + (b - 1)^2)\rho^2)]. \quad (3.23)$$

This function is the sum of two Gaussian modes: a quiescent mode centered at  $(0, 0)$  and an extreme mode centered at  $(1, 1)$ . The pdf is further controlled by two parameters:  $\gamma$  and  $\rho$ .

The parameter  $\gamma$  controls the weight of the extreme mode relative to the quiescent mode. When  $\alpha = 0$ , the extreme mode vanishes, and when  $\gamma = 1$ , the two modes contribute equal weight to the total pdf.

The parameter  $\rho$  controls the spacing of the modes, relative to the radius of the two modes. When  $\rho = 1$ , the standard deviation of each hump is  $\sqrt{2}$ , which is equal to the center spacing. When  $\rho > 1$ , the modes shrink in radius, increasing the relative distance between them.

Figure 3-9 shows the pdf for representative values of  $\rho$ .

Figure 3-9 shows the existence of a ‘knuckle’ in the shape of the QRS surface, near  $q = 0.1$ . This shape is due to the scale separation of the underlying pdf, and  $q$  corresponds to the small value of  $\gamma$  (0.05) that controls the mass distribution between the modes.

Figure3-9 also shows a knuckle, near  $q \approx 0.8$ , though it is less pronounced. Again, this is due to the underlying scale separation. However, because scale separation

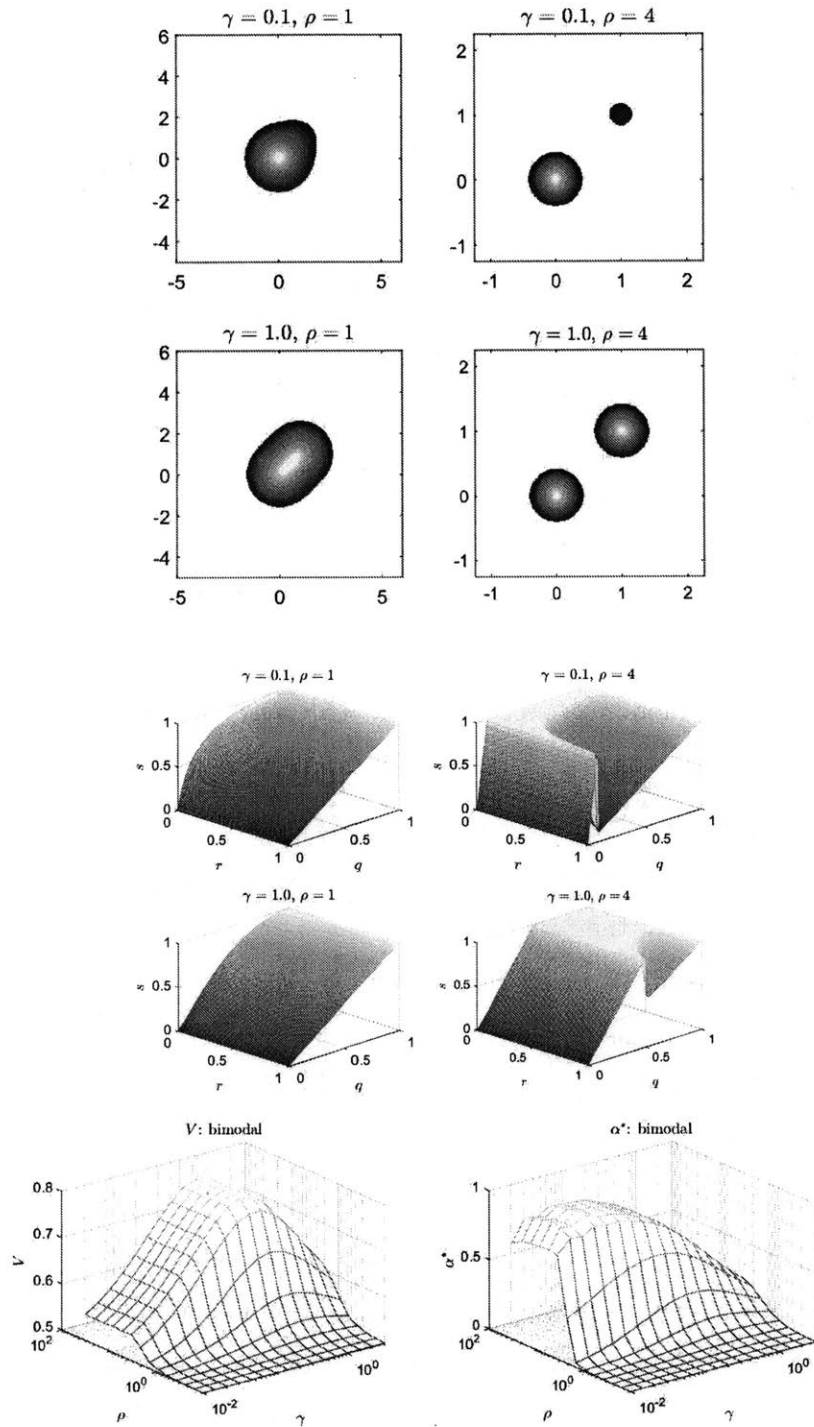


Figure 3-9: a) Joint pdf plots of the bimodal scenario for various parameters. b) Corresponding QRS plots. c, d) QRS metrics: c)  $V$ , d)  $\alpha^*$ .

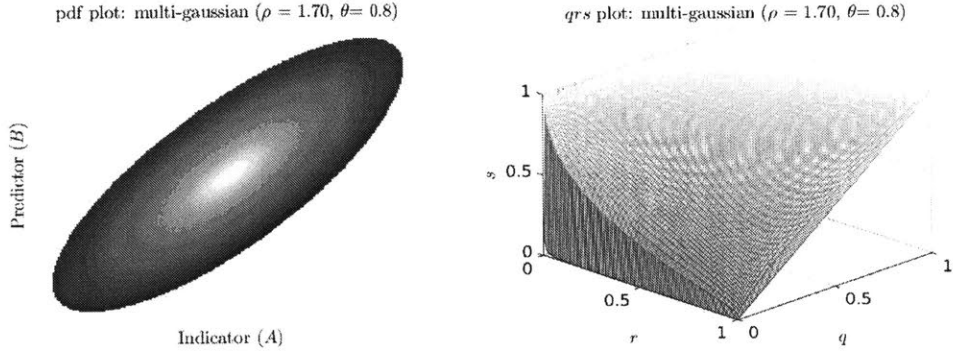


Figure 3-10: Sample a) pdf plot and b) QRS surface for the multivariate Gaussian scenario.

happens at a relatively low threshold  $\hat{a}$ , the feature is smaller. Alternatively, because the coinflip predictor is already fairly precise at  $q = 0.8$ , there is little ‘room’ for a good extreme event predictor to be better.

The plots in figure 3-9 show the various summary statistics of the bimodal scenario as a function of  $\gamma$  and  $\rho$ .  $V$  (figure 3-9 a) is largest when  $\rho$  is large (scale separation) and  $\gamma$  is medium (equal distribution of mass).

Unlike  $V$ , the knuckle metrics peak at *small*  $\gamma$ . This matches intuition from looking at the QRS plots in figure 3-9: the knuckle is larger when it occurs at low extreme event rate.

### 3.6.2 Multivariate Gaussian Predictor

The multivariate gaussian scenario is described by a pdf of the form

$$f_{ab}(a, b; \rho, \theta) = \frac{1}{\beta} \exp[-(\cos \theta a + \sin \theta b)^2 \frac{1}{\rho^2} - (\sin \theta a - \cos \theta b)^2 \rho^2]. \quad (3.24)$$

where  $\beta$  is a normalization factor,  $\rho^2$  is the ratio of length of the principal axis to the perpendicular axis, and  $\theta$  is the angle between the principal axis and the  $a$  axis.

Figure 3-10 a) shows a sample pdf. When  $\theta = 0$  or  $\theta = \frac{\pi}{2}$ , there is no linear (or higher) correlation between the indicator and predictor, and in both cases the

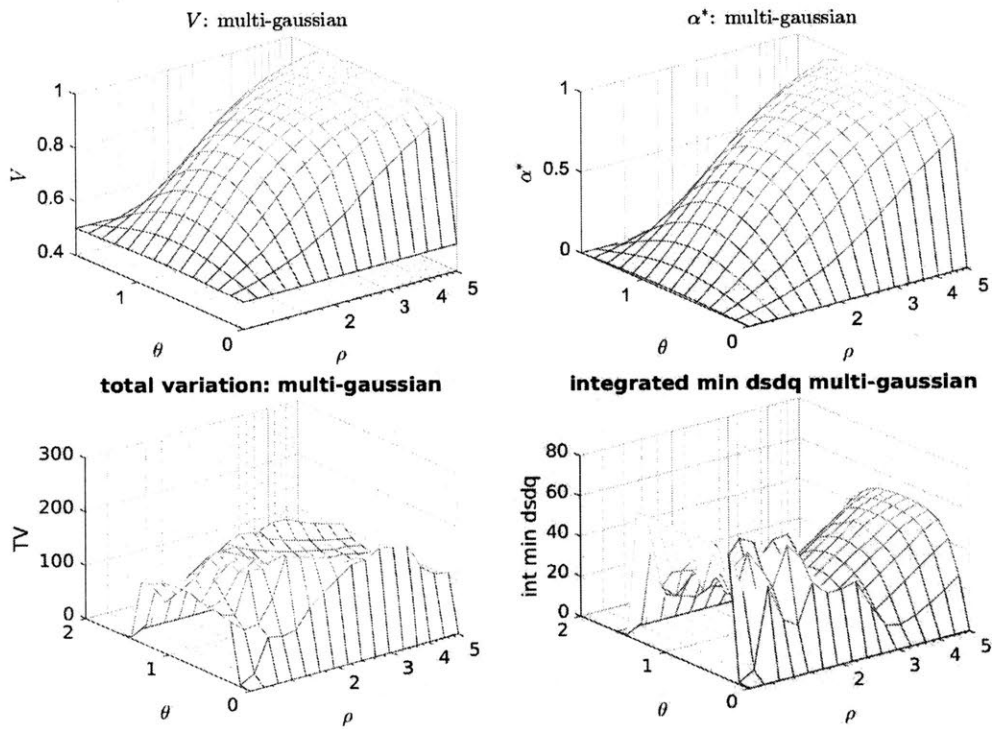


Figure 3-11: Sample summary statistics for the multivariate Gaussian scenario. a)  $V$ , b)  $\alpha^*$ , c)  $\eta_2$ , d)  $\eta_1$

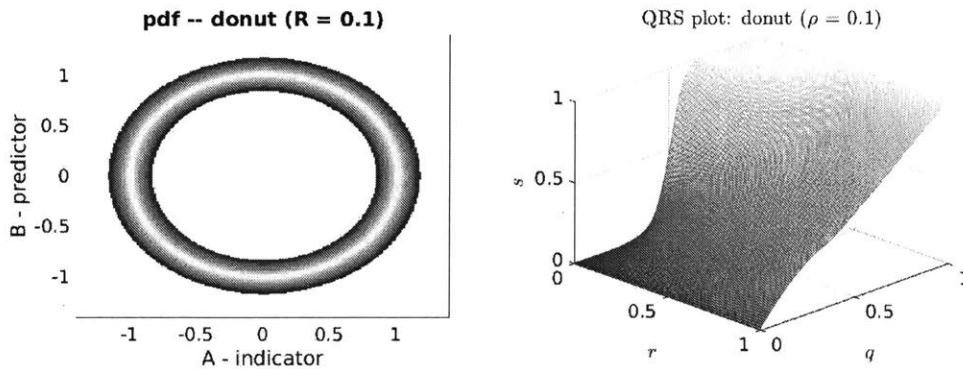


Figure 3-12: Sample a) pdf plot and b) QRS surface for the donut scenario,  $R = 0.1$ .

predictor is merely a (monotonically rescaled) coinflip predictor. However, when  $\theta$  is near  $\frac{\pi}{4}$ , and  $\rho > 1$ , there *is* a (linear) correlation between  $a$  and  $b$ . In this regime, we see the QRS surface (3-10 b) begin to swell, and the  $V$  (3-11 a) increase.

Unlike the bimodal scenario, there is no distinct scale separation. This appears in the QRS plot (3-10 b) as an absent knuckle. This missing knuckle also ‘appears’ in both  $\eta_2$  (3-10 c) and  $\eta_1$  (3-10 d) as a noisy constant background caused by numerical differentiation.

### 3.6.3 Donut Predictor

The donut scenario is described by a pdf of the form

$$f_{ab}(a, b; R) = \frac{1}{\beta} \sqrt{a^2 + b^2} \exp\left(-\frac{1 - a^2 - b^2}{R^2}\right). \quad (3.25)$$

where  $\beta$  is a normalization factor, and  $R$  is the thickness of the donut. The donut is designed so that there is some kind of significant relationship between  $a$  and  $b$ , but that relationship cannot be easily described as a linear correlation.

The shape of the QRS plot for the donut scenario (figure 3-12 b) is worth commenting on. For a small extreme event rate ( $q \lesssim 0.5$ ), the precision is actually *less* than the extreme event rate. This is because both large and small values of the predictor correspond non-extreme values of the indicator, while the most extreme events ( $a \approx 1$ ) correspond to intermediate values of the predictor.

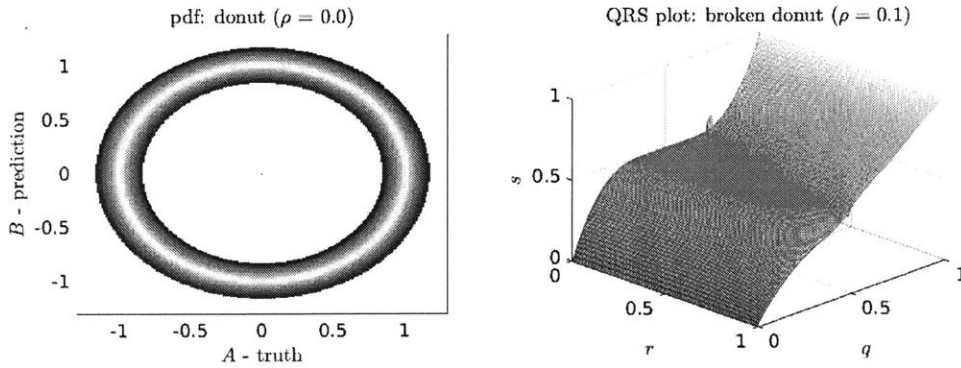


Figure 3-13: Sample a) pdf plot and b) QRS plot for the nonmonotonically fixed donut scenario.  $R = 0.1$

That is to say, the conditional probability  $P(a > \hat{a}|b)$  decreases with more extreme values of  $b$  in both directions. While there *is* some value of the predictor that corresponds to the most extreme events (and also the most quiescent events), it cannot ‘isolated’ by some simple monotonic transformation.

This is all to say that, despite the high mutual information between  $a$  and  $b$ ,  $b$  does not predict extreme events well, in the sense described by precision, and recall.

One way to turn the ‘bad’ predictor  $b$  into a better predictor  $b^*$  is to apply some kind of a non-monotonic transformation to it. Here, the (negative) absolute value of  $b$  appears to be closely related to the indicator. The resulting binary classification isn’t *precise* (at least, not for low ferocity) because there is a second ‘horn,’ but the predictor has become better than a coinflip for small extreme event rates.

This transformation wasn’t enough on its own, however. For large extreme event rates,  $b^*$  again has  $s < q$ . This is perhaps less important than the same issue for low extreme event rates, but it still suggests that  $b^*$  is a poor predictor of ‘extreme’ events which are not very extreme at all.

A quick look at the histogram (figure 3-13 a) identifies the problem. When the threshold for extreme events is very high ( $\hat{a} = 0.9$ , say), all of the extreme events have large  $b^*$  values. However, when the threshold for extreme events is very low ( $\hat{a} = -0.5$ , say), more of the extreme events have small (more negative)  $b^*$  values. Somewhere (around  $\hat{a} = 0$ , corresponding to  $q = 0.5$ ), the best test for extremity

swapped from high values of  $b^*$  to low values.

Often, this phenomenon might be corrected by choosing a better indicator. For instance, a pdf shaped like figure 3-13 a) might appear in the context of deep water waves if, instead of measuring the wave height, the signed maximum deviation from sea level were recorded. In this scenario, extreme events really correspond to both large *and small* values of the indicator, for which the QRS procedure is poorly suited.

Note that while the correct non-monotonic transformation was simple to choose in this case, it does not appear generically out of the QRS procedure. Non-monotonic rescaling of a potential predictor is too far removed for the ‘evaluation’ step, and more correctly belongs in the ‘selection’ step.

### 3.7 Appendix

For a predictor that generates a non-monotonic SR curve, the following construction builds a derivative monotonic SR frontier.

**Definition 3.7.1** (Precision Recall Frontier). For a fixed  $\hat{a}$ , the **precision recall frontier** (SR frontier) is the monotonic curve defined by

$$s^*(r) = \max_{r_0 \in [r, \infty)} (s(r_0)). \quad (3.26)$$

**Theorem 8** (Precision Recall Frontier Predictor). *The predictor corresponding to the regions traced out by definition 3.7.1 (the frontier) but not by definition 3.3.1 (the curve) can be achieved by the following construction.*

*For each  $r_2$  such that  $s(r_2) < s(r_1)$  for some  $r_1 > r_2$ , choose the values of  $\hat{a}$  and  $\hat{b}$  corresponding to the pair  $(s(r_1), r_1)$ . Without loss of generality, assume  $(s(r_1), r_1)$  is a local maximum of the SR-Curve (equation 3.7). Create a new binary classification with the following rule: all  $(a, b)$  are assigned to the quadrant appropriate to the thresholds  $\hat{a}$  and  $\hat{b}$  except that predicted extreme events with probability  $\gamma$  are reclassified as predicted quiescent events, true positives to false negatives and false positives to true negatives.*



In order to achieve  $r_2$  with a local maximum at  $r_1$  the fraction of events to reclassify  $\gamma$  is given by  $\gamma = 1 - \frac{r_2}{r_1}$

*Proof.* In order to show that the frontier exists and has the necessary properties, it suffices to show that each point on the frontier can be reached by some (modified) binary classification.

Precision is given by the ratio of successfully predicted extreme events to all predicted extreme events. If equal proportions are taken from the true positive quadrant and the false positive quadrant, the ratio is unchanged.

Recall is given by the ratio of successfully predicted extreme events to true extreme events,  $r_1 = \frac{n}{n+m}$ . If a fraction  $\gamma = 1 - \frac{r_2}{r_1}$  of the true positives are reclassified as false negatives, the new recall is given by

$$\frac{(1 - \gamma)n}{n + m} = \frac{r_2}{r_1} \frac{n}{n + m} = r_2. \quad (3.27)$$

□

The procedure easily extends to the entire QRS surface simply by constructing the described frontier for each constant  $q$  slice. The algebraic results (for  $V$ , or for  $\frac{\partial s}{\partial q}$ ) are greatly complicated, however.

# Chapter 4

## Machine Learning Paradigm

### 4.1 Problem Overview

The prediction problem so far sketched can best be formulated in the following way: “Given a dynamical system with (1) a particular observable exhibiting extreme events and (2) a set of other observable features, what function over (2) best predicts (1)?”

This is a search problem—in particular, an optimization problem. Because we can only just ‘best predict’ in terms of data (stored trajectories), this problem is a natural fit for a machine learning approach.

In order to apply machine learning to this problem, we need to make choices in four components areas. These components are:

- Hypothesis Class
- Training Data
- Objective Function
- Optimization Algorithm

Each choice is only weakly coupled to the others, so we will investigate each in turn.

## 4.2 Hypothesis Class

Extreme event prediction is conceptually similar to binary classification: at some point  $(x, t)$ , we want to know if there will be an extreme event near  $(x, t + \tau)$  by polling a predictor function  $B(x, t)$ . The goal of the learning task is to pick the best predictor  $B$ , called a *hypothesis*.

However, the space of all possible functions is prohibitively large. For a practical problem, we restrict our attention to some set of hypotheses  $\mathcal{B}$ , called the *hypothesis class*.

PAC learnability depends intimately on the choice of hypothesis class. In general, there are three considerations:

First is the bias/overfitting tradeoff. A very large hypothesis class may better fit the data, but it may also overfit to the training data noise. Conversely, a small hypothesis will generalize better, but may not be powerful enough to capture the important features in the training data.

Second, a larger hypothesis class represents a higher dimensional space, which is more difficult to search through. Especially in the case of rare events, evaluating the loss function for a given hypothesis may be time consuming, so exhaustive search through a large hypothesis class will be quickly overtaken by the curse of dimensionality.

Third, the hypothesis class may encode certain prior information about the problem. We may know a priori that certain kinds of predictors are very unlikely to be any good. Therefore, if we remove those terrible predictors from the hypothesis class (or judiciously choose  $\mathcal{B}$  to avoid them), we might spend less time testing bad hypotheses.

For dynamic systems with a physical interpretation, the choice of hypothesis class will likely represent the major application of physical intuition and domain knowledge.

## 4.3 Training Data

### 4.3.1 Overview of Issues

Initially, whether is drawn from simulation or observation, it is captured in the form of trajectories  $u(t)$ .

For instance, one possibility is that stock market data may be captured in the form of a vector time series, whether each vector component is one member of the NASDAQ exchange. In other case, the raw data may take the form of a simulation of fluid in a box, where each element of a time series is a 2D grid of vector fluid velocities.

In order to train our predictor, we need *training data*. A typical training data set  $\mathcal{D}$  is a list of ordered pairs  $(b, a)$  of the form prediction-truth.

However, in order to construct their pairs from trajectory data, we need to answer three **correspondence** problems:

When we make a prediction at time  $t_B$ , we mean that the predictor function may take as input information about the state at time  $t_B$  (and possible also at time  $t < t_B$ ). This prediction, made at time  $t_B$ , corresponds to a potential extreme event at time  $t_A$ . The first question to ask is, how are  $t_B$  and  $t_A$  related? A related question is ‘given that extreme events may be spread out in time, how ought predictions be spread out in time?’

Further, as in the case of the fluid simulation, extreme events may be spatially located. Similar questions arise as in the temporal case, but with the additional complication that extreme events and typical predictor functions both have associated spatial spreads, and their spreads may have quite different scales.

Finally, after the problems of correspondence are worked out, there is the question of how many training points should be kept. Ought there be one training point for each true extreme event (and then, some points correspond to quiescent events)? Or, should there be one training point for each space-time grid point, no matter how ‘big’ (in space or in time) the extreme event is?

We can break down the choice of correspondence relationship into three parts:

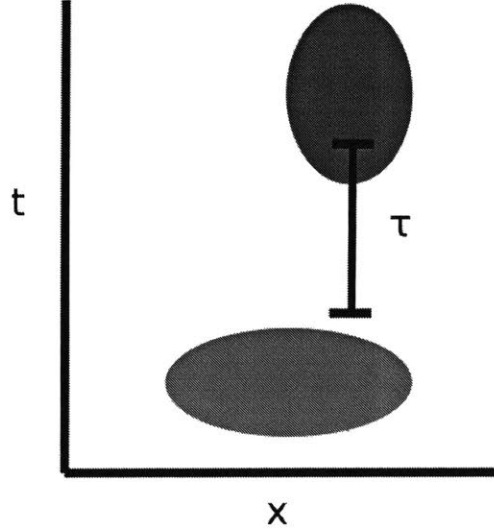


Figure 4-1: Cartoon representation of `strict time-lag` correspondence rule.

- temporal correspondence
- spatial correspondence
- sampling measure

### 4.3.2 Temporal Correspondence

In our problem, we are interested in *forecasting*, which means that we would like our predictor  $B$  to make use of only information that is (causally) prior to the extreme event at  $\mathbf{x}'$ .

This suggests another simple rule, `strict time-lag`. Under this rule, for a given  $\mathbf{x}$ ,  $\mathbf{x}'$  is chosen so that the spatial coordinate is identical, while the time coordinate is lagged by some interval  $\tau$ . Under this rule, the question becomes “what observable measurement  $B \in \mathcal{B}$  best approximates the extremeness-of-event at the same location  $\tau$  time later?”

This is a better rule, but there are two problems associated with its practical interpretation:

**First**, a positive prediction at  $\mathbf{x}$  may not correspond to an extreme event  $\tau$  later,

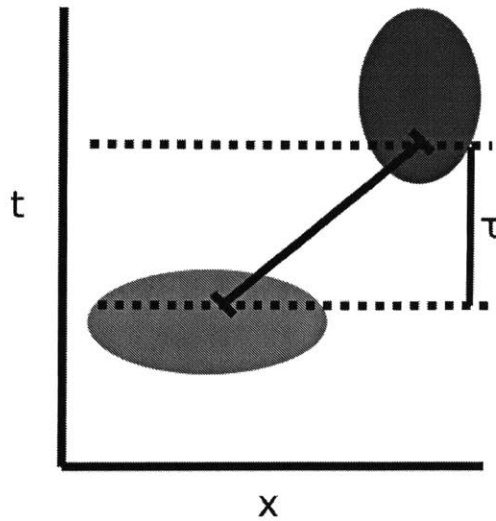


Figure 4-2: Cartoon representation of `space max` correspondence rule.

but to one  $\tau + \epsilon$  later. The `strict time-lag` rule will call the prediction a false positive even though there *is* a real extreme event coming later, and there *was* a ‘forecasted’ forewarning.

**Second**, a positive prediction at  $\mathbf{x}$  may correspond to a prediction slightly displaced in space.

The following rules for time-matching have been presented:

- `identity rule`
- `strict time-lag`

### 4.3.3 Spatial Correspondence

The requirements for data-matching in space differ slightly from those in time. First, there is no spatial forecasting: there is no meaningful sense of ‘ahead’ or ‘behind’ along a spatial axis. Second, most reasonable hypothesis classes are spatially distributed, while it is easy to imagine hypothesis classes that only sample the simulation data at a single time instant. Third, the spatial spread of predictors is variable (for instance, wavelets with variable length scale  $L$ ), while the temporal spread for forecasting is a mostly-fixed  $\tau$ .

The simplest rule is the **spatial identity rule**: that is, a prediction at  $x$  can only correspond to a (later) extreme event exactly at  $x$ . This is likely to lead to many **false positives** because  $\frac{\partial B}{\partial x}$  is likely to be smaller than  $\frac{\partial A}{\partial x}$  near extreme events.

A slight variation of the **identity rule** can be imagined if the simulation data includes a transport velocity  $u$ . If a prediction at  $(x, t)$  is advected to become a later extreme event at  $(x + u\tau, t + \tau)$ , then we can use the **advection rule** to make this correspondance. Like the **identity rule**, each prediction at  $\mathbf{x}$  corresponds to a single unique extreme event at  $\mathbf{x}'$ , but there is a spatial as well as temporal displacement.

One way to dodge the spatial issue is to somehow ‘integrate it out.’ In this approach, forecasting is *really* a problem about time series, not fields. That means, if we had a way to recast  $B(\mathbf{x})$  as  $B^*(t)$ , then we could avoid the whole issue of asking about whether a spatially imprecise prediction was correct or not.

The rule **space max** is simple:  $B^*(t) = \max_x B(x, t)$  and  $A^*(t) = \max_x A(x, t)$ . This is justified because extreme events are also rare. It is unlikely (in some sense that depends on the spatial extent of the problem and the true extreme event rate) that there will be two extreme events close together in time. Thus, a prediction at time  $t$  surely corresponds to a unique extreme event at time  $t + \tau$ .

Unlike along the temporal axis, in the spatial domain there is no problem of forecasting. For this reason **full neighborhood** is an excellent strategy for assembling predictor-indicator pairs. One remaining issue is the size of the neighborhood, while may depend on  $L$ .

The following rules for space-matching have been presented

- **identity rule**
  
- **advection**
  
- **space max**
  
- **full neighborhood**

### 4.3.4 Sampling Measure

Some of the previous correspondence strategies, such as `full neighborhood`, have raised the question about how many data points should be selected, and where. For instance, `full neighborhood` can be formulated both as “for every prediction, pick the most extreme event in its neighborhood” as well as “for every event, pick the most extreme prediction in its neighborhood.”

For bijective spatial and temporal correspondences such as `asidentity rule` and `strict time-lag`, sampling every such pair preserves the distribution of both  $a$  and  $b$  from the simulation data to the training data. This perfect translation is the measure rule `exact-pair`.

These two framings naturally lead to two choices of sampling measure rule: `every-prediction` and `every-event`.

In `every-prediction`, we make one training data point for every  $\mathbf{x}$  in the simulation domain. Further, every prediction is represented once. However, different predictions (made at points  $\mathbf{x}_1$  and  $\mathbf{x}_2$ ) may correspond to the same extreme event (at point  $\mathbf{x}'$ ). This is because there are many different places that a given extreme event could logically be predicted from.

The `every-prediction` measure rule is a natural setting for precision and accurately defining false positives, because it samples training data points according to the measure of the predictions. This rule guarantees that the distribution of  $b$  in the training data accurately represents the distribution of  $b$  in the simulation data.

In `every-event`, in contrast, we pick every event and look for what prediction ought correspond to it. Under this measure rule, two different (extreme) events may correspond to the same prediction. However, because extreme events are rare, it is unlikely that two different extreme events will point to the same prediction (under `full neighborhood`, say). Instead, a certain extreme event will point to its best prediction, but so will all the non-extreme events near to it.

This rule guarantees that the distribution of  $a$  in the training data accurately represents the distribution of  $a$  in the simulation data. It is a natural setting for



recall and accurately defining false negatives.

When combined with a correspondence rule like `full neighborhood`, there is a readily apparent problem with these two measure rules. The measure rule `every-prediction` guarantees the marginal distribution of  $b$ , at the expense of inflating the number of  $(a, b)$  pairs that point to a particular (extreme)  $a$ . Conversely, `every-event` guarantees the marginal distribution of  $a$  but over represents extreme  $b$ .

The following rules for sampling measure have been presented

- `exact-pair`
- `every-prediction`
- `every-event`

#### 4.3.5 Summary

The choice of how to turn trajectory data time series into training data pairs is not a straightforward one. In many cases, it is difficult to find a rule that satisfies both conceptual clarity and technical clarity.

## 4.4 Objective Function

The Empirical Risk Minimization (ERM) paradigm works by searching the hypothesis class  $\mathcal{B}$  for the hypothesis  $B^*$  that optimizes some loss function  $L[B]$  over a given set of training data  $\mathcal{D}$ . To this end, we must choose a loss function that closely accords with our intuition of what describes a good predictor.

The typical binary classification task minimizes the `total error rate` ( $T$ ), which is defined as

$$T = \frac{\# \text{ correctly classified}}{\# \text{ total}}. \quad (4.1)$$

This error metric is poorly suited for the extreme event prediction problem for two reasons:

**First**, total error rate is unsuited for unbalanced data. Extreme events are usually associated with extremely unbalanced data sets. This manifests in two ways. First, even a naive predictor may achieve  $> 99\%$  accuracy, simple because always predicting “not extreme” is usually correct. Second, resampling the data (for instance, to balance the number of extreme and not-extreme training points) may widely change the total error rate, which in turn may change the optimal predictor.

**Second**, this error metric is unsuited for strength-of-confidence measurement. It has no ability to distinguish between confidently classified points and and unconfidently classified points. This is particularly important if we expect our predictor to make many mistakes.

The first objection may be resolved by using balanced error metrics, such as the **balanced error rate** (the arithmetic mean of the sensitivity and the specificity) or the **F score**, given by the harmonic mean of the precision and the recall.

The second objection may be resolved by the use of Receiver Operating Characteristic (ROC) Curves. These curves represent graphically how different predictor thresholds give different tradeoffs between false positives and false negatives. The quantities developed in part 3 are designed to be particularly appropriate in the context of extreme events.

## 4.5 Optimization Algorithm

### 4.5.1 Choice of Algorithm

Because  $\mathcal{B}$  may be a high dimensional space, and because evaluating  $L[B]$  may be costly, it is important to develop an efficient strategy to find the optimal  $B$ .

This problem has three core properties that will inform our choice:

**First**, computing the loss function  $L[B]$  is extremely expensive. We should like a search algorithm that trades more overhead for fewer function calls.

**Second**, the loss function probably does not have an analytic derivative. This means that computing a local gradient is expensive, requires  $O(d)$  additional function

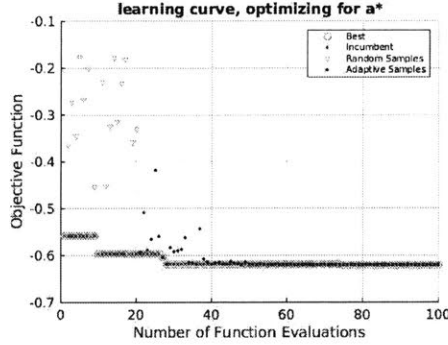


Figure 4-3: Sample learning curve for surrogate optimization. Note the change near  $n = 20$  from pseudo-random samples to adaptive samples.

calls.

**Third**, the loss function *may* be stochastic. Because the training data is generated from simulation, we always have the choice of simulating more data. If it improves our optimization, we might choose a different training data set for different function evaluations.

Because of points one and two, stochastic gradient descent (SGD) may not be a good fit for this problem. SGD, touched upon in section 2.7.2, is the gold standard for big data-based optimization, where it leverages the identity

$$\mathbb{E}[\nabla L] = \nabla \mathbb{E}[L]. \quad (4.2)$$

in order to efficiently travel down the gradient in an expected sense [12].

Instead, we will need to use one of gradient-free approaches discussed in section 2.7.3. Surrogate optimization is the best fit for this problem. The per-evaluation overhead introduced by the surrogate fitting is small compared to the expensive objective evaluations, and the exploration/exploitation balance is a natural fit for expensive black-box optimization.

## 4.5.2 Parametrization of Search Space

Surrogate optimization searches through a rectangular search space corresponding to the direct product of one bounded interval for each parameter. When the actual

problem has a different space, this can create a problem. There are three general possibilities: penalty terms, specialty symmetry solves, and nonlinear rescalings. Each will be presented briefly.

A **penalty term** is a term  $\Omega$  added to the objective function  $L^* = L^o + \Omega$ , defined as

$$\Omega(\mathbf{B}) = \begin{cases} 0 & \mathbf{B} \text{ permissible} \\ b & \text{otherwise} \end{cases} \quad (4.3)$$

In regions of the rectangle where the correct solution is disallowed, a large term is added to the objective to ‘force’ the optimizer to look elsewhere.

In order to avoid needless and expensive computations, this check should be performed *before* actual evaluation of the objective, (and  $b$  set suitable large) to avoid extra evaluations.

This method has the advantage of quick and simple implementation, for any shaped space that can be described by a nonlinear inequality. Its disadvantage is that the finite size of the radial basis functions causes search to be biased away from the *boundaries* of the permissible search space. Especially in high dimensional spaces, this may cause the optimizer to avoid large parts of the permissible region.

A *symmetry solver* is a custom implementation of the surrogate search that takes advantage of certain even symmetries in the search space. If the predictors  $B_1$  and  $B_2$  are two parametrizations of the same predictor, then time can be freed by not evaluating both. Further, the surrogate fit can be improves by enforcing this symmetry condition on the radial basis function surrogate as well.

The obvious disadvantage of a custom implementation is time to code and debug, and the absence of proprietary technology in commercially developed optimization packages.

Finally, a non-rectangular search space may be **rescaled** so that it fits into a rectangle. A (unordered) list of sizes  $L_1, L_2, \dots, L_k$  may be expressed as a (sorted) list of signed differences  $\delta_1, \delta_2, \dots, \delta_{k-1}$  and an initial length  $L_0$ . This breaks the symmetry of the interchange between  $L_i$  and  $L_j$ . A further potential nonlinear rescaling

transforms the resulting triangular region into a rectangle.

## 4.6 Potential Questions

There are two ways to compare the optimal predictor from different ML passes:

**First**, we can compare the test error. That is, we can compute the optimal predictor on a new set of testing data, which helps control for overfitting.

**Second**, we can compare the hypotheses directly, by measuring some kind of kernel  $k(B_1, B_2)$ . If two different ML passes give similar optimal predictors, then we might say that the optimal predictors is good, because it is not sensitive to algorithm particulars.

# Chapter 5

## Application I –

## Majda-McLaughlin-Tabak Model

### 5.1 Model Overview

The Majda-McLaughlin-Tabak (MMT) model is a 1D nonlinear model of deep water wave dispersion first introduced by Majda et al in [2], and since studied in the context of weak turbulence and intermittent extreme events [17], [9], [67].

The governing equation is given by

$$iu_t = |\partial_x|^\alpha u + \lambda |\partial_x|^{-\frac{\beta}{4}} (|\partial_x|^{-\frac{\beta}{4}} u)^2 |\partial_x|^{-\frac{\beta}{4}} u + iDu, \quad (5.1)$$

and the Fourier transform

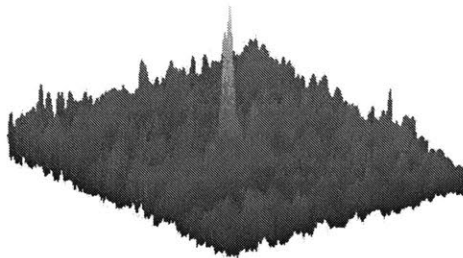


Figure 5-1: Sample plot of one simulated realization of the MMT model near an extreme event.

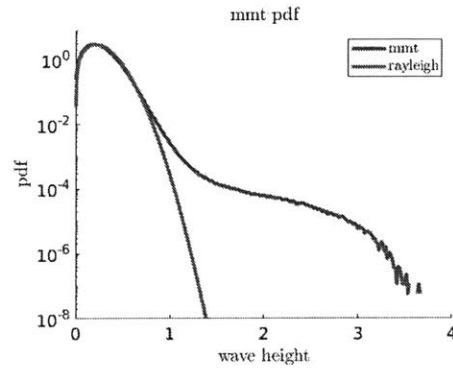


Figure 5-2: Probability density function of the MMR wave height. Rayleigh distribution overlaid for comparison—note the ‘long-tail’ extending from  $x \approx 1.5$  to  $x \approx 3.5$ .

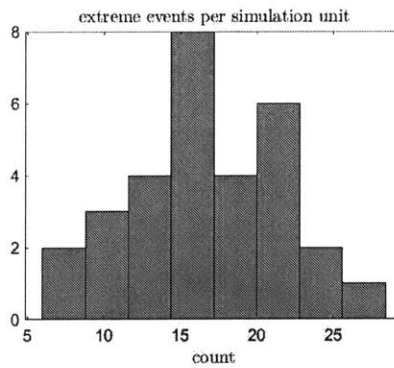


Figure 5-3: Histogram of the number of extreme events per simulation run.

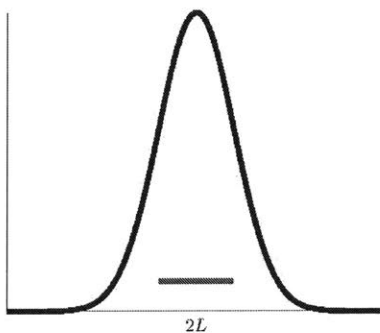


Figure 5-4: Cartoon representation of Gabor frame.

$$|\partial_x|^\alpha \widehat{u}(k) = |k|^\alpha \widehat{u}(k).$$

This work will follow Cousins and Sapsis [67], who simulated the MMT equation to validate a model for nonlinear wave collapse and extreme event prediction. The domain has spatial extent  $2\pi$ , discretized into 8192 points, and temporal extent 150, discretized into 6000 points for integration purposes. The parameters are chosen so that  $\lambda = -4$  (focusing case),  $\alpha = \frac{1}{2}$  (deep water case), and  $\beta = 0$ . The operator  $Du$  is a selective Laplacian designed to model dissipation at short scales, i.e., from wave breaking. The first 1000 points are discarded to avoid transients due to random initial conditions; no forcing term is included and the simulations represent free decay.

We chose the truth-indicating function

$$A(x_0, t_0) = |u(x_0, t_0)|, \tag{5.2}$$

which is a measure of wave group amplitude. With this identification, extreme events are large-amplitude wave groups, with  $A \gtrsim 1.5$  as seen in figure 5-3.



## 5.2 Method

### 5.2.1 Hypothesis Space

For the MMT problem, we expect that energy density at certain length scales might be related to extreme events. For this reason, we choose as our hypothesis class linear combinations of  $k$  zero-order Gabor coefficients of variable length scales [24].

The Gabor coefficient is given by the inner product of an element of the Gabor Frame with the data. The Gabor functions are given by

$$h_L(x; n) = \exp\left(-\frac{x^2}{2L^2}\right) \exp\frac{i2\pi nx}{L}, \quad (5.3)$$

and the zero-order case is achieved when  $n = 0$ . These functions can be conceptualized as localized wavelets.

The Gaussian kernel 5.3 is taken from Cousins and Sapsis [67], where the  $n = 0$  mode was found to predict energy transfer in the extreme event collapse.

The simplest such hypothesis class,  $\mathcal{B}_1$  is a one parameter space given by  $B[L_1] = G[L_1]$ . More complicated classes can be constructed via linear combinations: the hypothesis class  $\mathcal{B}_n$  contains elements  $B[\alpha_1, \dots, L_1] = \sum_k^n \alpha_k G[L_k]$ , which has  $2n - 1$  free parameters after an overall scaling constant is removed.

While the many equivalent parametrizations of  $\mathbf{w} = \{\alpha, L_1, L_2\}$  do not give different predictors, they do lead to differently ‘shaped’ hypothesis spaces  $\mathcal{B}_2$ . For instance, swapping  $L_1$  and  $L_2$  (followed by an appropriate transformation of  $\alpha$ ) gives an identical predictor. Searching the space  $\mathcal{B}_2$  is made easier by proactively removing this sort of symmetry.

Additionally, another step to regularize the hypothesis class is to replace the wavelength  $L$ , which could conceivably span 3 orders of magnitude, with the log wavelength  $\log(L)$ . This better aligns the measure of the parameter space with both the conceptual measure and metric appropriate for a radial basis function approximation of the parameter space.

## 5.2.2 Binning

The MMT model has both spatially and temporally located extreme events. Following section 4.3, we use the following rules to convert simulation data to training data: `strict time-lag`, `space max`, and `exact-pair`.

## 5.2.3 Objective Function

Following 4.4, we will use four objective functions:

- total accuracy, representing a default machine learning metric
- $F_1$  score, representing a standard metric for unbalanced data
- `alpha star`, which measures scale separation
- `VUS`, which measures classification across all thresholds

When we compare the ROC metrics to the standard binary classification metrics, we will use a threshold  $\hat{a} = 1.5$ , which roughly corresponds to the edge of the quiescent range.

## 5.2.4 Optimization Loop

We use Matlab's implementation of surrogate search. Except where otherwise noted, we terminated optimization after 4 hours of runtime on consumer desktop hardware, which represents  $\approx 100$  function evaluations when calculating from 1 unit of simulation data.

# 5.3 Results

## 5.3.1 Features of Optimal Metric

For each objective function, the optimal predictor has a similar form:

- a short length scale component

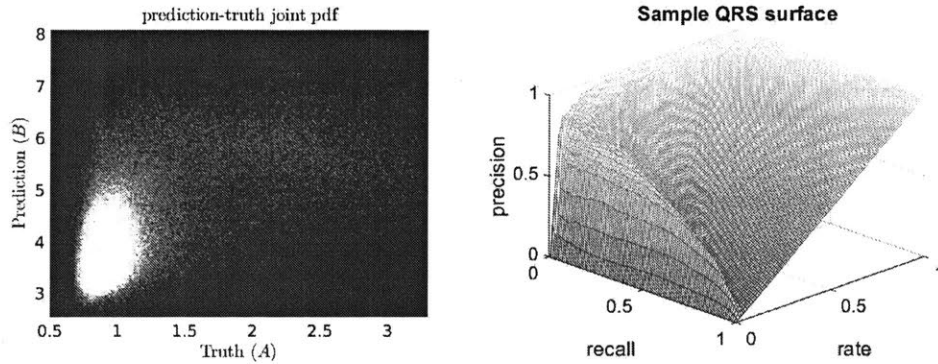


Figure 5-5: a) Sample prediction-truth joint pdf for a good MMT predictor. b) Corresponding ROC surface plot.

- a long length scale component
- a amplitude weighting greatly favoring the short component

This breakdown has a simple physical interpretation: in order for an extreme event to occur, there must be sufficient background energy to draw up (long length scale), and also enough localized “seed” energy which will begin the collapse. This interpretation agrees with previous work by Cousins and Sapsis [67].

The joint prediction-truth pdf (figure 5-5 a) has one major features: the greater density of events by far is in the lower left corner, the true negatives. Outside of this region, events *seem* widely spread, and a straightforward visual inspection is difficult.

The associated ROC surface plot (figure 5-5 b) offers a little more insight. The knuckle feature near  $r = 0.05$  suggests that there *is* some scale separation going on; and that a predictor can do better than chance at predictor which extreme events will exceed that threshold.

### 5.3.2 Comparison of Optimal Predictors

#### Comparison between Objective Functions

The different choices of objective function, of course, lead to different optimal parameters, as shown in figure 5-6. Three of the objectives,  $f_1$ ,  $vus$ , and  $\alpha^*$  result in

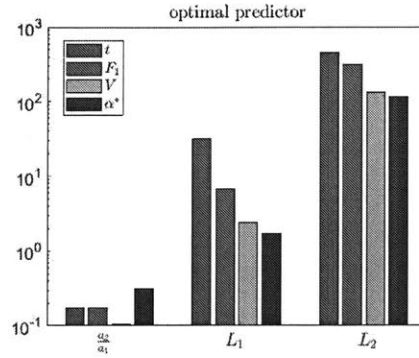


Figure 5-6: Optimal predictor parameters for each objective function. Note that total accuracy is very different than the others.

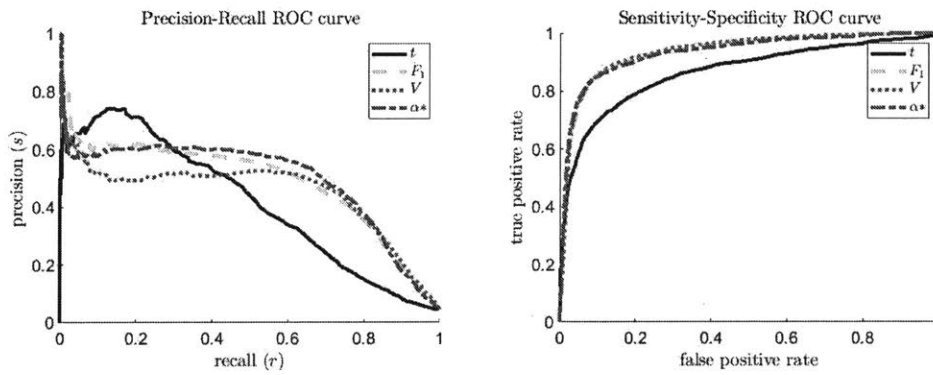


Figure 5-7: Receiver Operating Characteristic Curve comparisons of optimal predictors calculated via different objectives. a) precision-recall curve b) sensitivity-specificity curve.

similar parameter values, while total accuracy  $t$  is quite different—generally, longer length scales.

A natural question is, how do the predictions (of the different optimal predictors) differ, and which is better?

Figure 5-7 shows two sets of ROC curves for the different optimal predictors. In the sensitivity-specificity setting (figure 5-7 b), all predictors are excellent, but total accuracy is less excellent all around. The precision-recall setting (figure 5-7 a) is more useful, and shows that the total-accuracy-optimized predictor can achieve slightly greater precision at very low recall tolerances, but otherwise performs more poorly.

This can be interpreted as follows: because of the particular balance between false positives and false negatives that total accuracy makes, and the widely unbalanced data sets that contain far more quiescent events than extreme events, total accuracy overvalues precision (avoiding false positives). The other metrics, which are designed to achieve more balance, sacrifice some precision for more consistent recall (fewer false negatives).

### Other Hypothesis Spaces

For this investigation, we ran numerical experiments on a two-vector Gabor coefficient space of predictors, which had three adjustable parameters. We can imagine other, more complicated spaces which might contain better predictors.

Exploratory investigations of three-vector predictors (with five adjustable parameters) almost invariably collapsed onto two-vector solutions: that is, given  $L_1$ ,  $L_2$ , and  $L_3$  for the the three-vector optimal predictor, either  $L_1 \approx L_2$  or  $L_2 \approx L_3$ . This suggests that the two length-scale interpretation of the optimal predictors given in subsection 5.3.2 is not just a necessary artifact of hypothesis space dimension.

Examining optimal predictors shows that they rarely exceed 0.5 precision by a significant amount. This 50% accuracy rate reflects an approximate temporal symmetry of the MMT extreme event mechanism: focusing and defocusing energy distributions look very similar when the time rate of change is ignored.

As a result of this symmetry, predictors from the two-vector Gabor space cannot reliably distinguish between ‘before extreme event’ progenitors (which will go on to become extreme events) and ‘after extreme event progenitors’ (which will not).

A more complicated hypothesis space, perhaps which compared Gabor coefficients between different time steps, or which uses the Gabor coefficients calculated directed from data finite differences, would likely be able to make this distinction, leading to predictors that achieve much higher precision.

### 5.3.3 Learning Rate

#### Learning Rate: Time

Figure 5-8 shows the effect of changing run-time on optimal parameters. In each case, the optimal predictor is highly noisy when the surrogate optimizer is emphasizing ‘exploration,’ and then converges as the optimizer switches over to ‘exploitation.’

In every case, the optimal parameter *variance* never drops to zero. This remnant variance represents the contribution from sampling error in the training data; because there are so few extreme events in each simulation, swapping out between simulations will result in slightly different optimal predictors.

The total accuracy has relatively high variance in the length scale parameters, while  $F_1$  has high variance in the amplitude ratio parameter.  $\alpha^*$  has the overall lowest intrinsic variances.

#### Learning Rate: Data

In addition to changing the optimization time, the size of the training data set may be varied. However, unlike time, there is a subtlety here.

First, the *total optimization time* may be kept constant. Increasing the size of the data set increases the objective evaluation time in direct proportion, so doubly the size of the training data will halve the number of evaluations.

Second, the *number of objective evaluations* may be kept constant. Again, because of objective evaluation time, doubly the size of the training data will cause the total

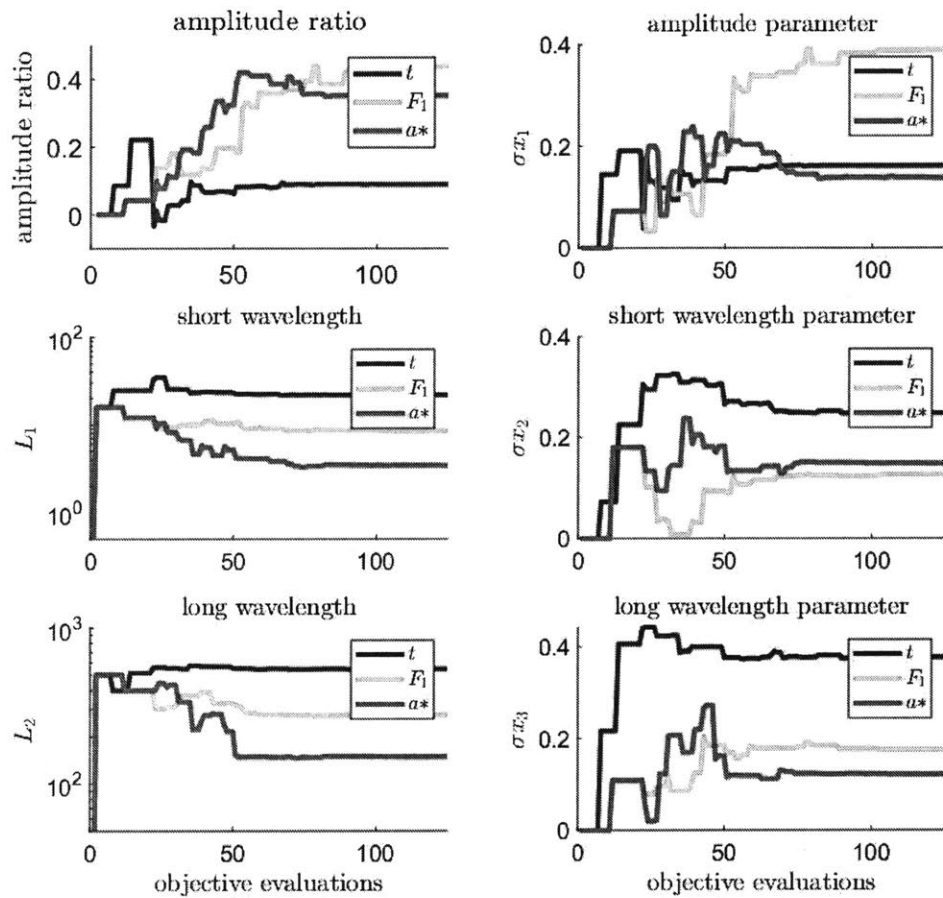


Figure 5-8: Learning curves as a function of time. a) optimal parameters, b) parameter variance.

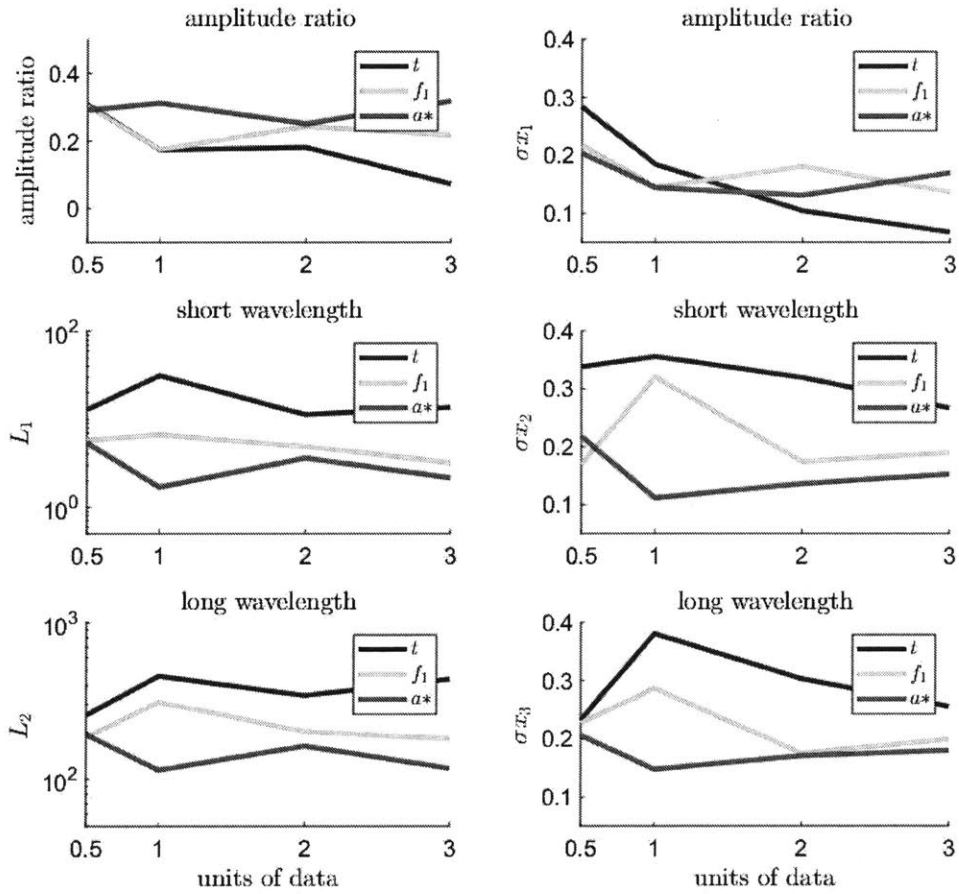


Figure 5-9: Learning curves as a function of data. a) optimal parameters, b) parameter variance.



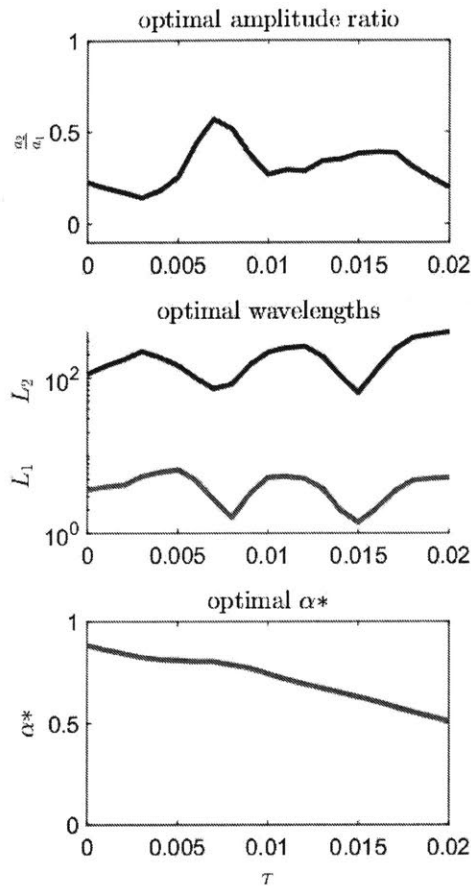


Figure 5-10: a,b ) Optimal predictor parameters as a function of  $\tau$ . c) Optimal  $\alpha^*$  as a function of  $\tau$

optimization run-time to double.

In order to capture this trade-off, figure 5-9 uses the first model, that is, given fixed run-time, what is the best balance between speed and data?  $\alpha^*$  generally finds the optimal balance at 1 simulation unit and consistently shows less variance in the optimal parameters.

### 5.3.4 Effects of Time Gap

In the previous numerical experiments, a fixed time gap  $\tau = 0.015$  has been used to represent a suitable prediction time scale: long enough for significant wave evolution,

short enough that good predictions are better than blind chance.

Figure 5-10 shows the optimal predictor parameters associated with the  $\alpha^*$  objective function and other choices of  $\tau$ . Generally, the trend is to weigh the long length scale component slightly more heavily as  $\tau$  increases, but overall the trend is small.

Additionally, there appears to be a dramatic feature near  $\tau = 0.007$ , where both optimal length scales precipitously drop, accompanied by a corresponding change in the amplitude ratio. This dramatic graph obscures the fact that the predictor parameters are interdependent—the increase of the amplitude ratio partially counteracts the effects of the shorter length scales.

Unfortunately, ‘predictor parameter values’ is not always a good ‘predictor similarity’ metric. A better metric would examine how closely the prediction made by two candidates across a given data corpus align, but care would have to be taken to avoid numerous fluctuations in the true negative mode from dominating.



# Chapter 6

## Application II – Kolmogorov Flow Model

### 6.1 Model Overview

The Kolmogorov flow is a solution to the forced Navier Stokes problem on 2D periodic domain. Above  $Re \approx 35$ , the solution is unstable, and there are intermittent bursts of energy dissipation.

The Navier Stokes equations (pressure-velocity form), defined over some domain  $\Omega$ , are given by

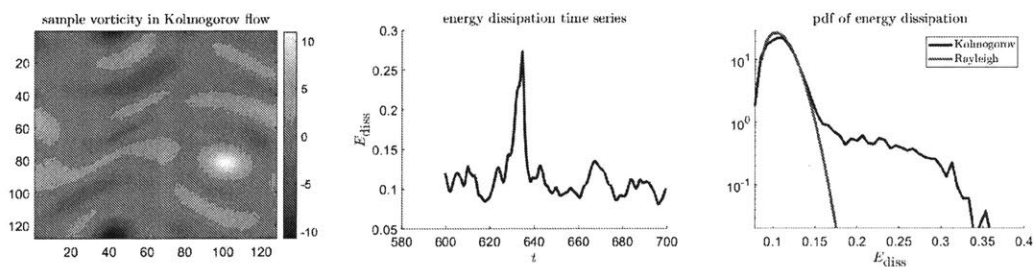


Figure 6-1: Descriptive plots for the Kolmogorov Flow. a) Sample realization of the the vorticity. b) Time series of energy dissipation near an extreme event. c) PDF of the energy dissipation

$$\begin{aligned}\partial_t u &= -u \cdot \nabla u - \nabla p + \nu \Delta u + f \\ \nabla \cdot u &= 0,\end{aligned}\tag{6.1}$$

where  $u$  is the (vector valued) fluid velocity field,  $p$  is the (scalar valued) pressure field,  $\nu$  is the dimensionless viscosity (inversely related to the famous Reynolds number) and  $f$  is some forcing term.

In the Kolmogorov Flow model, the forcing is a monochromatic time invariant field given by

$$f(\mathbf{x}) = \sin(\mathbf{k}_y \cdot x) \hat{e}_1,\tag{6.2}$$

where  $\mathbf{k}_y = (0, 4)$  is the wavenumber of the forcing field and  $\hat{e}_1 = (1, 0)$  is a unit vector perpendicular to  $\mathbf{k}_y$ .

The intermittent bursting phenomena associated with the Kolmogorov flow for large enough Reynolds numbers ( $\text{Re} \gtrsim 35$ ) are captured by the energy dissipation rate, given by

$$D(u) = \frac{\nu}{|\Omega|} \int_{\Omega} |\nabla u|^2 dx.\tag{6.3}$$

Farazmand and Sapsis [43] studied the Kolmogorov Flow and determined that extreme values in certain Fourier modes of the flow correspond to increased energy influx, which predict later bursts of energy dissipation. In particular, the energy input rate

$$I(u) = \frac{1}{|\Omega|} \int_{\Omega} u \cdot f dx\tag{6.4}$$

reliably reaches a peak shortly before the energy dissipation.

### 6.1.1 Method

#### Hypothesis Space

A natural global set of predictors are the coefficients associated with low- $k$  2D Fourier modes,  $b_{\mathbf{k}}$ . We'll also consider arbitrary linear combinations of these coefficients, that is, predictors given by

$$B = \sum_{\mathbf{k}} \gamma_{\mathbf{k}} b_{\mathbf{k}}. \quad (6.5)$$

#### Other Choices

The remaining methodological choices closely track those discussed in 5.2, with the exception that Fourier predictors are global, so there is no spatial binning of any sort.

### 6.1.2 Results

#### Overview

Figure 6-2 shows the prediction quality of the single-coefficient predictors according to different metrics. By every metric, the Fourier coefficient with wavenumber  $k = (0, 4)$  is an especially good predictor.

Some other Fourier modes are also identified as strong potential predictors: in particular, modes with  $k_y - k_x \geq 4$ . That these modes predict bursts of dissipation is consistent with a theory that describes energy dissipation as resulting from an energy cascade through smaller and smaller length scales.

#### Effects of Time Gap

Figure 6-3 plots the six most significant  $\gamma_{\mathbf{k}}$  from equation 6.5 in order to show the effects of changing  $\tau$  on the composition of the optimal predictors.

The most important component from the combined predictor is the  $(0, 4)$  mode, and its weighting factor is always positive. Other Fourier modes, such as  $(3, 0)$  have a negative weighting factor which means they are inversely correlated with bursts of

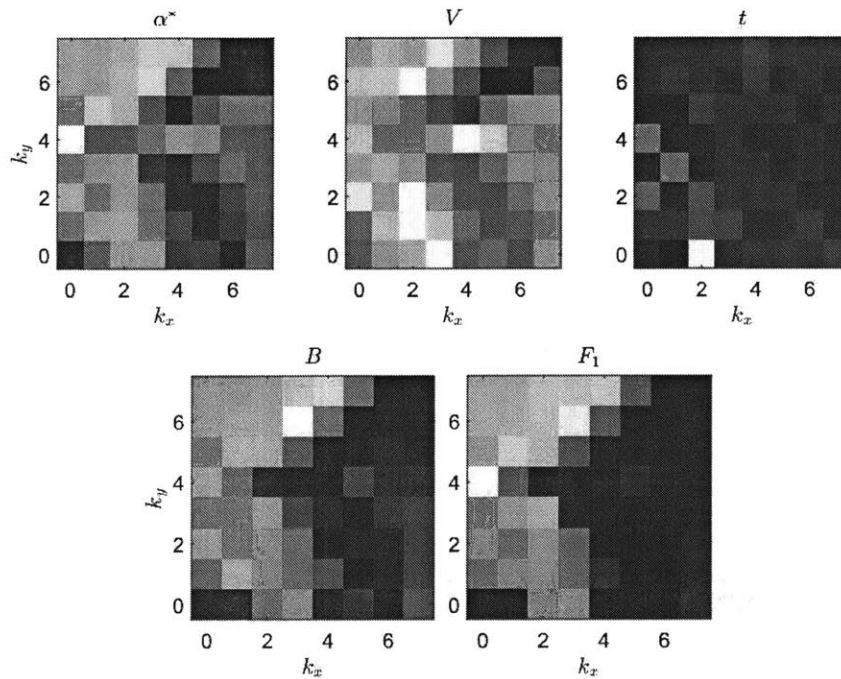


Figure 6-2: Plots of single coefficient predictor quality for different wavenumbers and objectives. a)  $\alpha^*$ , b) Volume under the surface, c) total accuracy, d) balanced accuracy, e)  $F_1$  score. Note the consistent peak at  $(0, 4)$ , which is resolved best by  $\alpha^*$  and  $F_1$ .

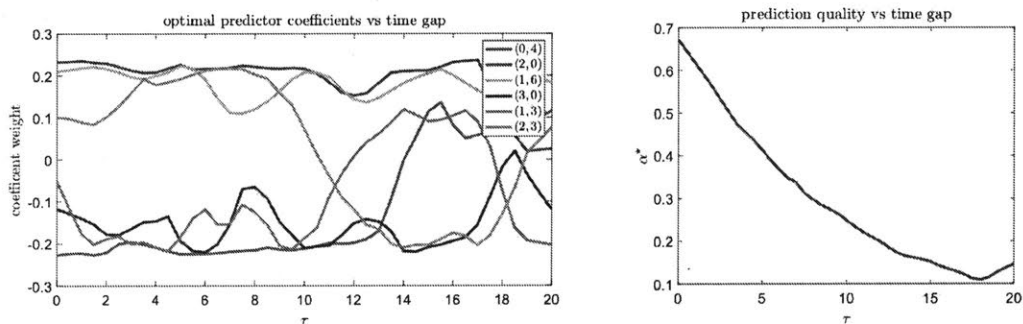


Figure 6-3: a) Composition of optimal predictor, in terms of Fourier modes, and b) quality of optimal predictor, each as a function of prediction gap  $\tau$ .

extreme dissipation. Due to the consistent downward trend in prediction quality as  $\tau$  increases, trends in the data past  $\tau \approx 15$  are less likely to be meaningful.





# Chapter 7

## Conclusions and Further Work

### 7.1 Conclusion

In this thesis paper, we have shown a method for optimizing extreme event prediction in an equation free manner. We have shown how the QRS surface construction allows for a geometric interpretation of scale separation, and naturally leads to the metric  $\alpha^*$  which is well suited to this problem. We compared  $\alpha^*$  to other metrics in two models of extreme events, where we showed that  $\alpha^*$  selects for qualitatively better predictions than the total accuracy, and has superior optimization properties as compared to  $F_1$ .

### 7.2 Further Work

This thesis examined in depth a way to construct a prediction metric suited to the problem of extreme event prediction. However, we spent comparably less time on the questions of selecting good hypothesis classes, and the binning procedure to go from trajectory data to training pairs. While we believe both these questions are highly problem dependent, any attempt to apply the machine learning paradigm to a related problem must address these issues either explicitly or otherwise.



# Bibliography

- [1] *On entropy approximation for Gaussian mixture random vectors*, 2008.
- [2] E. G. Tabak A. J. Majda, D. W. McLaughlin. A one-dimensional model for dispersive wave turbulence. *Journal of Nonlinear Science*, 7(1):9–44, 2 1997.
- [3] T. Schreiber A. Kaiser. Information transfer in continuous processes. *Physica D*, pages 43–62, 2002.
- [4] Henry D.I. Abarbanel. *Analysis of Observed Chaotic Data*. Springer, 1996.
- [5] Kuniyiko Kaneko Akiko Kashiwagi, Itaru Urabe and Tetsuya Yomo. Adaptive response of a gene network to environmental changes by fitness-induced attractor selection. *PLoS ONE*, 1:1–10, 2006.
- [6] Andras Sobester Alexander Forrester and Andy Keane. *Engineering design via surrogate modelling: a practical guide*. Wiley, 2008.
- [7] Harald Stogbauer Alexander Kraskov and Peter Grassberger. Estimating mutual information. *Physical Review E*, 69, 6 2004.
- [8] Jean-Luc Blanc Annick Lesne and Laurant Pezard. Entropy estimation of very short symbolic sequences. *PHYSICAL REVIEW E*, 79:046208–1–10, 4 2009.
- [9] Laura Biven Benno Rumpf. Weak turbulence and collapses in the majdaaŠm-claughlinãŠtabak equation: Fluxes in wavenumber and in amplitude space. *Physica D*, (204):188–203, 4 2005.
- [10] Fischer Black and Myron Scholes. The pricing of options and corporate liabilities. *The Journal of Political Economoy*, 81(3):637–654, 1973.
- [11] Ronald R. Coifman Ioannis G. Kevrekidis Boaz Nadler, Stephane Lafon. Diffusion maps, spectral clustering, and eigenfunctions of fokker-planck operators. In J.C. Platt B. Scholkopf and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, 2006.
- [12] LÃon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT 2010*, pages 177–186, 2010.
- [13] Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

- [14] Kristian Tylen Andreas Roepstorff Jacob F. Sherson Dan Monster, Riccardo Fusaroli. Causal inference from noisy time-series data – testing the convergent cross-mapping algorithm in the presence of noise and external influence. *Future Generation Computer Systems*, 2016.
- [15] Georges A. Darbellay and Igor Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Interactions on Information Theory*, 45(4):215–220, 5 1999.
- [16] David W. McLaughlin David Cai, Andrew J. Majda and Esteban G. Tabak. Spectral bifurcations in dispersive wave turbulence. *Proceedings of the National Academy of Sciences of the United States of America*, 96(25):14216–14221, 1999.
- [17] David W. McLaughlin Esteban G. Tabak David Cai, Andrew J. Majda. Dispersive wave turbulence in one dimension. *Physica D*, (152-153):551–572, 2001.
- [18] Thomas G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 160(7):1895–1923, 2 2017.
- [19] Gretchen G. Moisen Elizabeth A. Freeman. A comparison of the performance of threshold criteria for binary classification in terms of predicted prevalence and kappa. *Science Direct*, 217:48–58, 2008.
- [20] Mohamad Farazmand and Themistoklis P. Sapsis. A variational approach to probing extreme events in turbulent dynamical systems. *Science Advances*, (3):215–220, 9 2017.
- [21] Francesco Fedele and M. Aziz Tayfun. On nonlinear wave groups and crest statistics. *J. Fluid Mech.*, 620:221–239, 2009.
- [22] Andrew M. Fraser and Harry L. Swinney. Independent coordinates for strange attractors from mutual information. *Physical Review A*, 33(2):1134–1140, 1986.
- [23] Mark I. Freidlin and Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*. 2012.
- [24] D. Gabor. Theory of communication. *Journal of Institution of Electrical Engineers*, 93(3):429–457, 1946.
- [25] Hao Ye Chih-hao Hsieh Ethan Doyle Michael Fogarty George Sugihara, Robert May and Stephan Munch. Detecting causality in complex ecosystems. *Science*, 338:496–500, 2012.
- [26] C. W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):44–438, 8 1969.
- [27] Haibo He and Edwardo A. Garcia. Learning from imbalanced data. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 21(9):1263–1284, 2009.

- [28] Desmond J. Higham. An algorithmic introduction to numerical simulation of stochastic differential equations. *SIAM Review*, 43(3):525–546, 2001.
- [29] Shie Mannor Huan Xu, Constantine Caramanis. Robustness and regularization of support vector machines. *Journal of Machine Learning Research*, 10:1485–1510, 2009.
- [30] Markos A. Katsoulakis and Petr Plechac. Information-theoretic tools for parametrized coarse-graining of non-equilibrium extended systems. *The Journal of Chemical Physics*, 139:074115–14, 8 2013.
- [31] Roshan Kumari and Saurabh Kr. Srivastava. Machine learning: A review on binary classification. *International Journal of Computer Applications*, 160(7):11–15, 2 2017.
- [32] Aida C. G. Verdugo Lazo and Pushpa N. Rathie. On the entropy of continuous probability distributions. *IEEE Transactions of Information Theory*, 24(1):120–122, 1 1978.
- [33] Annick LESNE. Shannon entropy: a rigorous mathematical notion at the crossroads between probability, information theory, dynamical systems and statistical physics. *Developments of the concepts of Randomness, Statistic, and Probability*, 24, 6 2014.
- [34] Fan Li. Modelling the stock market using a multi-scale approach. Master’s thesis, University of Leicester, 2017.
- [35] Wentian Li. Mutual information functions versus correlation functions. *Journal of Statistical Physics*, 60(5/6):823–837, 3 1980.
- [36] Adam B. Barrett Lionel Barnett and Anil K. Seth. Granger causality and transfer entropy are equivalent for gaussian variables. *Physical Review Letters*, 103:238701–4, 12 2009.
- [37] Andrea Caponnetto-Michele Piana Lorenzo Rosasco, Ernesto De Vito and Alessandro Verri. Are loss function all the same. *Neural Computation*, 16(5):1063–1076, 2004.
- [38] Guy Lapalme Marina Sokolova. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45:427–437, 2009.
- [39] James J. Riley Martin R. Maxey. Equations of motion for a small rigid sphere in a nonuniform flow. *Physics of Fluids*, 26(4):883–889, 1982.
- [40] James M. McCracken and Robert S. Weigel. Convergent cross-mapping and pairwise asymmetric inference. *Physical Review E*, 90, 12 2014.

- [41] Chris Tofallis Michael Leznik. Estimating invariant principal components using diagonal regression.
- [42] Mustafa A. Mohamad and Themistoklis P. Sapsis. Probabilistic description of extreme events in intermittently unstable dynamical systems excited by correlated stochastic processes. *SIAM/ASA J. Uncertainty Quantification*, 3:709–736, 8 2015.
- [43] Themistoklis P. Sapsis Mohammad Farazmand. Reduced-order prediction of rogue waves in two-dimensional deep-water waves. *Preprint*, Preprint:Preprint, Preprint.
- [44] Will Cousins Mustafa A. Mohamad and Themistoklis P. Sapsis. A probabilistic decomposition-synthesis method for the quantification of rare events due to internal instabilities. *Journal of Computational Physics*, 322:288–308, 6 2016.
- [45] Liam Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15:1191–1253, 2003.
- [46] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2:559–572, 1901.
- [47] Stephane Lafon Ronald R. Coifman. Diffusion maps. *Applied and Computational Harmonic Analysis*, (21):5–30, 2006.
- [48] Kenny Daily S. Joshua Swamidass, Chloe-Agathe Azencott and Pierre Baldi. A roc stronger than roc: measuring, visualizing and optimizing early retrieval. *BIOINFORMATICS*, 26(10):1348–1356, 2010.
- [49] Alison L. Marsden Sethuraman Sankaran, Charles Audet. A method for stochastic constrained optimization using derivative-free surrogate pattern search and collocation. *Journal of Computational Physics*, 229:4664–4682, 2010.
- [50] Nathan Srebro Shai Shalev-Shwartz, Yoram Singer. Pegasos: Primal estimated sub-gradient solver for svm. In *Proceedings of the 24th International Conference on Machine Learning*, 2007.
- [51] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.
- [52] Abdullah Z. Sheikh. Non-normality of market returns, 2009.
- [53] Marco Vannucci Silvia Cateni, Valentina Colla. A method for resampling imbalanced datasets in binary classification tasks for real-world problems. *Neurocomputing*, 135:32–41, 2014.
- [54] Marc Rehmsmeier Takaya Saito. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE*, 10(3):1–21, 3 2015.

- [55] Floris Takens. Detecting strange attractors in turbulence. *Lecture Notes in Mathematics*, 898, 1981.
- [56] Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.
- [57] Joy. A. Thomas Thomas M. Cover. *Elements of Information Theory*. Wiley Interscience, second edition edition, 2006.
- [58] Dimitirios Giannakis Tyrus Berry and John Harlim. Nonparametric forecasting of low-dimensional dynamical systems. *Physical Review E*, 91(032915):1–7, 2015.
- [59] John Harlim Tyrus Berry. Iterated diffusion maps for feature identification. *Applied and Computational Harmonic Analysis*, pages 1–36, 2016.
- [60] John Harlim Tyrus Berry. Variable bandwidth diffusion kernels. *Applied and Computational Harmonic Analysis*, 40:68–96, 2016.
- [61] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(6):265–280, 1971.
- [62] Liz Cheek Viv Bewick and Jonathan Ball. Statistics review 13: Receiver operating characteristic curves. *Critical Care*, 8(6):508–512, 2004.
- [63] Youssef Hamadi Vu Khac Ky, Claudia D’Ambrosio and Leo Liberti. Surrogate-based methods for black-box optimization. *International Transactions in Operational Research*, 2016.
- [64] Zhong Yi Wan and Themistoklis P. Sapsis. Machine learning the kinematics of spherical particles in fluid flows. *Journal of Fluid Mechanics*, 857:1–11, 2018.
- [65] Yilun Wang and Christine A. Shoemaker. A general stochastic algorithmic framework for minimizing expensive black box objective functions based on surrogate models and sensitivity analysis. *ArXiv*, 2014.
- [66] Norbert Wiener. *The Theory of Prediction*. McGraw-Hill, 1956.
- [67] Themistoklis P. Sapsis Will Cousins. Quantification and prediction of extreme events in a one-dimensional nonlinear dispersive wave model. *Physica D*, (280-281):48–58, 5 2014.
- [68] E. Zio and N. Pedroni. Estimation of the functional failure probability of a thermal-hydraulic passive system by subset simulation, journal = .