

CLARKSON UNIVERSITY

**Inference of networks of causal relationships using
Causation Entropy**

A Dissertation by

Warren M. Lord

Department of Mathematics

Clarkson Center for Complex Systems Science

Submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

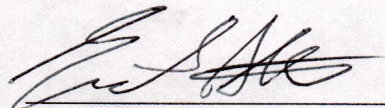
Mathematics

April 27, 2018

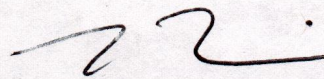
The committee below have examined the thesis dissertation entitled “**Inference of networks of causal relationships using Causation Entropy**” presented by **Warren M. Lord**, a candidate for the degree of **Doctor of Philosophy (Mathematics)**, and have certified that it is worthy of acceptance.

4/23/18

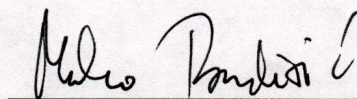
Date



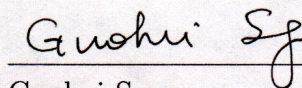
Erik Bollt (Advisor)



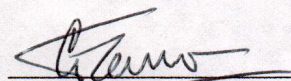
Jie Sun (Co-advisor)



Marko Budišić



Guohui Song



Christino Tamon

Abstract

Many interesting phenomena in nature and society are generated by the interactions of a large number of simpler components. Examples include global climate, the functioning of a living cell, and human cognition. In order to understand how the behaviors of individual components combine to create large scale emergent phenomena, it is important to understand the interactions between the components.

Defining and identifying the interactions is often a nontrivial problem. The thesis proposes three criteria for such a definition: 1) the relationships are predictive, 2) they do not depend on model-specific assumptions, and 3) the definition distinguishes between direct and indirect relationships.

These criteria lead to the consideration of information theory. A notable contribution of the thesis is a unification and generalization of differential entropy and Shannon entropy. This unification is made possible by an unconventional definition of Kullback-Leibler divergence which makes different assumptions from the definition traditionally used in information theory.

The consideration of information theory in a dynamical setting leads to notions of “information flow”, such as Transfer Entropy (TE). Causation Entropy (CSE) generalizes TE and satisfies the three criteria defining a causal relationship.

The optimal Causation Entropy algorithm (oCSE) is used with CSE to efficiently identify causal relationships. The application of oCSE to real world problems requires

estimating CSE from time series data. This thesis gives an introduction to estimation that is novel in that it is both purely measure-theoretic and synthesizes the principles of parametric and non-parametric statistics.

The thesis introduces geometric k-nearest neighbors estimators which are shown to outperform conventional k-nearest neighbors (knn) estimators in the task of estimating mutual information when the underlying dynamical systems are either dissipative or have multiple time scales.

The utility of oCSE and knn estimation is demonstrated with an application to insect swarming. The nodes are stochastic processes describing the trajectories of individual insects in a number of mating swarms in a laboratory environment. The results suggest that the insects often do not interact with their nearest spatial neighbors.

Acknowledgements

I would like to express my great appreciation to my advisor, Erik Bollt. His guidance, wisdom, and support have contributed greatly to my academic development. His diligent work with me over the past four years has helped me improve as a mathematician and set me up for a life of success. He has challenged me to work hard, and yet always injected a sense of humor and fun into our work together. I could not have hoped for a better advisor.

I would also like to thank Jie Sun, who has been a second advisor to me. In addition to his advice and encouragement I would like to thank him for always challenging me. His hard questions have forced me to think more deeply about our research and related mathematical concepts.

I am also grateful to Christino Tamon for the many hours of fascinating discussions in his office. From quantum computing, to Lie groups, to Haar measures, and machine learning, our chats have always been something to look forward to. I am especially grateful for his positive attitude and encouragement.

I would like to thank Marko Budišić and Guohui Song for their useful comments and suggestions on the thesis.

Finally, I want to thank my mother and father for their support and encouragement throughout my studies.

Contents

1	Introduction	1
2	Networks of causal relationships	6
2.1	Causality for scientists	6
2.1.1	Methods based on structural form	6
2.1.2	Predictive causality	10
2.2	Shannon Entropy	12
2.2.1	Some intuition about information	12
2.2.2	Shannon’s treatment of entropy	14
2.2.3	Alternative ways of thinking about entropy	18
2.3	Extending Shannon entropy to differential entropy	20
2.4	Absolute continuity, Kullback-Leibler divergence, and united frame- work for Shannon and differential entropies	27
2.4.1	Shannon Entropy and differential entropy as Kullback-Leibler divergence	28
2.4.2	Applications of differential entropy	35
2.4.3	Generalization	44
2.5	Mutual information	47
2.6	Transfer Entropy	55

2.7	Causation Entropy	57
2.8	The optimal Causation Entropy algorithm	60
2.9	Interpretation of the value of CSE as a real number	65
3	Data based inference of causal relationships	68
3.1	Background	69
3.1.1	The estimation problem as optimization of risk	70
3.1.2	Strategies	74
3.2	Nonparametric estimation of differential entropy and mutual information	82
3.2.1	Estimation of differential Entropy	82
3.2.2	Extension of differential entropy estimators to mutual information	87
3.3	A class of knn estimators that adapts to local geometry	91
3.3.1	Introduction	91
3.3.2	Method	95
3.3.3	The g-knn estimates for $H(X)$ and $I(X;Y)$	100
3.3.4	Examples	101
3.3.5	Discussion	110
3.4	Nonparametric Hypothesis Testing	113
4	Application to collective animal motion	116
4.1	Background	117
4.2	Experimental Methods	119
4.3	Inferring Insect Interactions using oCSE: Choice of variables, parameters, and conditioning	121
4.4	Results	125
5	Future directions	130

5.1	Application to brain function	131
5.1.1	Background	132
5.1.2	Causation Entropy and new data sets	133
5.2	Entropy and estimation	135
Appendices		141
A Probability primer		142
A.1	Random variables, distributions, and expectation	142
A.2	Sub- σ -algebras, and product measures	145
A.3	Absolute continuity, Radon Nikodym, and conditioning	150
A.4	Convergence of random variables	155
A.5	Jensen's Inequality	157
A.6	The Radon-Nikodym theorem	158

Chapter 1

Introduction

Many phenomena in nature, society, and technology seem to arise from the collective interactions of large sets of simpler components. For instance, schools of fish can synchronize their motions to avoid predators in a way that would not be possible by an individual fish [92]. Humans make social connections with other humans in ways that make it possible for ideas and viruses to spread rapidly across an entire society [13]. Human DNA consists of thousands of genes, many of which interact by slowing or speeding up the expression of other genes, in a way that gives rise to the function of the cell and allows the cell to adapt to changing conditions [83]. The human brain consists of over 10^{10} neuronal cells [56] (in addition to non-neuronal cells), which interact via synapses to give rise to problem solving and emotion. The brain also has mesoscopic structures that form anatomical units, which could also be thought of as nodes communicating via information pathways to solve problems and produce emotions.

The systems in each of these phenomena can be described as directed graphs (also called networks). Directed graphs can be defined as a set of vertices, \mathcal{V} together with a set of edges, $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. It is often the case that the nodes are easily identified. For

example, nodes might be fish in a school of fish, or genes in a cell. More generally, the nodes typically correspond to observables, or measurable quantities associated with a system.

The edges are more elusive. In general terms they represent direct causal relationships, but they are difficult to identify, since in many cases they do not correspond to a physical connection between the nodes. For example it is unclear what signal a fish sends out that causes another fish to respond. In other cases the mechanism is understood, but observing the interactions would be unfeasible. For instance, we know how humans share viruses but we often do not know who causes a particular person to get sick. We know that genes produce proteins that can interfere with or accelerate the rate that other genes produce proteins, but the interactions occur at the molecular level so that they cannot be directly observed. The neuronal network that forms the human brain has too many neurons to trace down every synapse between neurons. The information pathways between mesoscopic brain regions would likely vary from task to task and emotion to emotion, and therefore seem to reflect something more than the static physical structure of the brain.

Although it would be hard, if not impossible, to directly identify the edges in these examples, it is often easier to observe the system in motion. For instance, a set of cameras could be used to record the positions of the fish over time. A microarray could be inserted into a cell with thousands of sites that fluoresce when a protein, or a chemical associated with the production of that protein is abundant. Humans can be placed inside fMRI machines, or wear EEG equipment, that records activity at brain sites while they perform different tasks.

An edge in each of the examples might be evidenced by a statistical relationship between the states of variables over time. Simple examples of a statistical relationship are, “whenever brain region 1 becomes highly active, brain region 2 becomes more

active over the ensuing 2 seconds,” and, “when the amount of the protein produced by gene A increases there is usually a subsequent decrease in the amount of protein produced by gene B.” More complicated statistical relationships would be needed to ensure that the relationships that are inferred are direct relationships.

Statistical relationships may not represent valid causal relationships. In many of these cases no experiment could be performed to verify the inferred graph since direct manipulation of one of the variables would either be unethical or would change the behavior of the entire system. Instead, the inferred graphs could be validated by prediction. For instance using an inferred graph to successfully predict the future movement of a fish, or the entire school, would suggest that the relationships were indeed causal relationships. Therefore, the causal relationships that the researcher seeks might be called predictive statistical relationships.

The goal in each of these examples can be thought of as learning structure from observing dynamics. Thus, this problem can be seen as an inverse problem to the forward problem of describing the effect of network topology on dynamics. Being able to learn network topology from dynamics could be very important for the understanding of complex systems. Although there is no precise definition of a complex system, an important feature in addition to being a networked dynamical system, is the emergence of function, properties, or phenomena, that exist at macroscopic levels (for instance the entire system) but not evidenced by the dynamics at the microscopic levels (for instance the dynamics of individual nodes). In the above examples, the coordinated function of a school of fish to evade a predator, and cognition and consciousness of the human brain are considered emergent properties. It is possible that learning the topologies of the networks that form complex systems could shed light on the nature of emergence.

Chapter 2 concerns the mathematical definition of a causal relationship. The

chapter gives some background on different approaches to defining causal relationships, and concludes with an introduction to, and definition of, Causation Entropy (CSE) [124]. In addition to reviewing work that was done prior to the start of this thesis, this chapter also introduces a novel perspective on entropy. A generalization of the notion of Kullback-Leibler divergence, based on absolute continuity and the Radon-Nikodym derivative, leads to a generalization and unification of differential and Shannon entropy.

Chapter 3 concerns mathematical methods for using time series data to provide evidence for a causal relationship. This chapter serves two purposes. It gives an introduction to the theory of estimation that is novel in that it is both purely measure-theoretic and synthesizes the principles of parametric and non-parametric statistics. It also introduces a novel class of nonparametric estimators called geometric k-nearest neighbors (g-knn) estimators. It is shown that a problem with traditional applications of k-nearest neighbors estimation methods is that they can be very biased when the underlying dynamical system is dissipative or has multiple time scales. A particular (g-knn) estimator is developed which solves these problems. Much of the presentation of the new estimator is taken from Warren M Lord, Jie Sun, and Erik M Bollt, “Geometric k-nearest neighbor estimation of entropy and mutual information,” *Chaos: An Interdisciplinary Journal of Nonlinear Science* **28.3** (2018) [72].

Chapter 4 describes a novel application of CSE to collective animal motion. In this application the nodes are individual insects which are part of a larger swarm. At a macroscopic level the swarm stays together and keeps roughly the same shape, indicating the existence of communication channels among the individual insects. A causal edge from one insect to another indicates that the first insect is receiving information from the second and using it to adjust its flight path. The chapter shows how the insects’ positions are measured, turned into time series data, and used to

infer the networks that form the structure of the swarm. A novel feature of these networks is that they are dynamic, meaning that they evolve in time. Much of the presentation is taken from Warren M Lord, Jie Sun, Nicholas T Ouellette, and Erik M Bollt, “Inference of causal information flow in collective animal behavior,” *IEEE Transactions on Molecular, Biological and Multi-Scale Communications* **2.1** (2016) [73].

Chapter 5 discusses some ideas for future research. Although the estimator introduced in Ref. [72] shows great promise for the estimation of differential entropy and mutual information for low sample sizes, it may be possible to improve the asymptotic properties describing the behavior of the estimator as the sample size increases unboundedly. Another project is the inference of information pathways in the human brain. A final idea concerns the correspondence between the causal relationships inferred by two scientists watching the same phenomenon from different coordinate reference frames.

This thesis takes a probabilistic approach to the above questions. For instance, the nodes are assumed to be random variables in the measure-theoretic sense. This approach shifts the focus from individual data points in the state space to relationships between measurable functions that take values in that space. The approach results in a novel introduction to information theory in Ch. 2 and a unified approach to parametric and nonparametric estimation in Ch. 3. The necessary background in probability is covered in Appendix 1.

Chapter 2

Networks of causal relationships

2.1 Causality for scientists

Although the concept of causality is intuitive, and the language of causality is used in everyday speech, defining a framework for causality that is practical and useful to scientists studying complex systems is difficult. Two common frameworks, which are suitable for different types of scientific inquiry, are methods based on structural forms of systems of equations, and ones based on improvement in prediction.

2.1.1 Methods based on structural form

The framework of causality based on the structure of systems of equations is motivated by the idea that direct manipulation of the value of a causal variable will result in predictable changes in the variable that is being affected. A commonly cited method for determining causality in this sense has been devised by Judea Pearl [93]. Consider

a system of relationships between the variables in $\mathcal{V} = \{X_1, \dots, X_m\}$ defined by

$$X_1 = f_1(\mathcal{P}_1, U_1), \quad (2.1)$$

$$\vdots$$

$$X_i = f_i(\mathcal{P}_i, U_i), \quad (2.2)$$

$$\vdots$$

$$X_m = f_m(\mathcal{P}_m, U_m), \quad (2.3)$$

where $\mathcal{P}_i \subset \mathcal{V}$, and the U_i are unobserved variables, which introduce randomness into an otherwise deterministic set of relationships. The system defined by Eqs. (2.1) to (2.3) defines a directed graph with nodes \mathcal{V} and an edge $X_i \rightarrow X_j$ if $X_i \in \mathcal{P}_j$. Call such a graph a structural graph for a system of equations like Eqs. (2.1) to (2.3).

Pearl lays out a set of conditions on the system defined by Eqs. (2.1) to (2.3) under which the edges in the structural graph can be called causal. Among the conditions, the functions f_i are assumed to be “autonomous,” meaning that changing one of the equations by directly manipulating an X_i would not disrupt the other equations. An “intervention” in X_i is defined by replacing Eq. (2.2) with $X_i = x_i$ for a particular value of x_i , and treating X_i as an unvarying parameter in the other equations. If the system is autonomous then the structural graph of the system with Eq. (2.2) fixed by an intervention would be the subgraph of the original structural graph obtained by removing the node X_i . The directed graph defined by a system of equations that satisfy Pearl’s assumptions has the structure of a directed acyclic graph (DAG) between the observed and unobserved variables. Starting with some assumptions on the distributions of the unobserved variables, U_i , Pearl’s approach builds a directed acyclic graph in which the edges are causal relationships [93].

By including the effects of unobserved variables and updating their distributions

by directly manipulating the observed variables, Pearl’s method avoids the pitfalls of confounding variables. Although this framework for causation is popular in some areas of science, it has some features that make it unsuitable for finding causal relationships in complex systems research. One is that the focus on DAGs rules out feedback loops, which could be important to self-organization and emergence in animal or human interactions, gene networks, and human cognition. Another problem is that the system of interventions seems unfeasible in a high dimensional nonlinear dynamical environment. More generally, the variables in nonlinear dynamical systems can become inextricably intertwined over time, so that the conditions required to do Pearl’s causality analysis are unlikely to hold.

Pearl’s interventionist framework is not the only method based on the structure of a system of equations. Liang and Kleeman’s formalism considers the case where the static system, Eqs.(2.1) to (2.3), is replaced with a stochastic process [71, 106, 107].

Definition 2.1.1 (Stochastic process). A stochastic process is a collection of random variables taking values in the same state space, where the variables are indexed by a set, T , called the index set. When T has a linear ordering, such as when $T = \mathbb{Z}$, or $T = \mathbb{R}$, T is often called “time,” and the stochastic process is said to be indexed by time.

Since the applications depend on time series data, the index set will usually be assumed to be $T = \mathbb{N} = \{1, 2, \dots\}$, in which case the system is called a discrete time system. For the purposes of describing many systems T can be taken to be \mathbb{R} , and the system called a continuous time system. In this thesis the time index will be indicated with a subscript. For instance, $\{X_t\}_{t \in \mathbb{N}}$ is a stochastic process, and if there are multiple stochastic processes with the same time index, then they will be denoted with a superscript such as $\{X_t^1\}_{t \in \mathbb{N}}$. A stochastic process will sometimes be referred

to by removing the time subscript, so that $X = \{X_t\}_{t \in \mathbb{N}}$. Note that a collection of stochastic processes X^1, X^2, \dots, X^m indexed by \mathbb{N} can be equivalently thought of as a single stochastic process, (X^1, X^2, \dots, X^m) taking values in the Cartesian product of the sample spaces¹.

Liang and Kleeman [71, 106, 107] consider stochastic processes defined by systems of equations such as

$$X_t^i = f_i(A_{i1}X_{t-1}^1, A_{i2}X_{t-1}^2, \dots, A_{im}X_{t-1}^m, U_{t-1}), \quad (2.4)$$

where A_{ij} are fixed parameters and U_t is an independent stochastic process that represents noise. System (2.4) determines a structural graph where the nodes are the processes X^i , and there is an edge $X^j \rightarrow X^i$ if and only if $A_{ij} = 1$. The matrix $\{A_{ij}\}_{i,j=1,\dots,m}$ is called the adjacency matrix because it determines which nodes are “next to” each other in the structural graph. Liang and Kleeman also consider systems of differential equations and partial differential equations in which the structural graph can be defined in a similar fashion.

The Liang-Kleeman method identifies transfers of information between variables (see Sec. 2.6 for a more detailed explanation of information transfer). It defines these information transfers by comparing the behavior of the original system to the behavior after an intervention.

Many of the problems with the application of Pearl’s framework to complex system also apply to the Liang and Kleeman formalism. For instance fixing a variable in a stochastic process could dramatically alter the relationships between the remaining variables. As a simple illustration, fixing a variable in the Lorenz system destroys chaotic dynamics, often leading to fixed point behavior. It is therefore un-

¹Refs. [14, 109] give more examples and a good introduction to the theory of stochastic processes.

clear whether anything like the “autonomous” condition described by Pearl exists for complex systems. A further problem with the Liang and Kleeman formalism is that from a practical standpoint it seems to require *a priori* knowledge of the update rules, f_i^2 .

2.1.2 Predictive causality

An alternative notion of causality was introduced by Norbert Wiener in 1956 [19, 136], and later given a practical implementation by Clive Granger in 1969 [47]. This notion is defined by the two conditions:

1. The cause comes before the effect.
2. The cause and the effect share information³ that is not contained in any other available sources.

The method applies specifically to stochastic processes.

By the first of the Wiener-Granger conditions, if X_t^1 causes $X_{t'}^2$ then $t < t'$. The second of the conditions could have many interpretations. The condition is generally interpreted in a sense of predictive power. Qualitatively, if (X^1, \dots, X^m) is a stochastic process, then X^1 is said to G-cause X^2 if the best predictor of X_{t+1}^2 based on the set of variables $\{X_s^i\}_{s \leq t, i=1, \dots, m}$ is better than the best predictor of the same quantity but with the variables $\{X_s^1\}_{s \leq t}$ removed from the set of predictors.

As Clive Granger points out, these conditions do not necessarily imply a causal relationship between variables since there can always be unmeasured confounding variables. However, Granger also points out that if one restricts their attention to

²Pearl’s framework seems to avoid needing specific knowledge of f_i by modeling the distribution of the variables in \mathcal{U}_i , and using chain rules for conditional distributions, along with interventions, to find distributions of connected variables in the DAG.

³The term information here is meant colloquially, and not necessarily in the sense defined in Sec. 2.2.

what can be observed, then one can use Granger’s analysis to test for causal relationships within that set of observables, which provides one justification for the use of the word “causal” [48]. G-causality seems to capture what is meant by causality for many scientists who study time series data, and as long as the results are only interpreted to be “causal” within the scope of the observed variables, then there should not be any confusion in using causation-related terminology.

One of the assumptions that are often imposed in order to make computations more tractable is that the distributions of the stochastic processes (See Appendix 1) are stationary.

Definition 2.1.2 (Stationary distribution). A stochastic process, $\{X_t\}_{t \in T}$, has a stationary distribution if the distribution of X_t does not depend on t (See App. 1 for the definition of distribution). In other words, for all $t_1, t_2 \in T$, $\nu_{X_{t_1}} = \nu_{X_{t_2}}$.

Clive Granger built the first implementation of these conditions by interpreting the improved predictive power in terms of linear regression. Given a time lag, $\tau \in \mathbb{N}$, X^2 is said to G-cause X^1 if the removal of the X^2 term in the regression

$$X_{t+\tau}^1 = A_{11}X_t^1 + \overbrace{A_{12}X_t^2}^{\text{remove}} + \cdots + A_{1n}X_t^n + E_t^1 \quad (2.5)$$

increases the variance of the error term, E_t^1 . More generally, one can compare the variances of the errors after fitting the multivariate regressions

$$X_t^1 = A \cdot [X_{t-1}^1, X_{t-2}^1, \dots, X_{t-p_1}^1, X_{t-1}^3, \dots, X_{t-p_2}^3, \dots, X_{t-p_2}^m]^T + E_t \quad (2.6)$$

$$X_t^1 = A \cdot [X_{t-1}^1, X_{t-2}^1, \dots, X_{t-p_1}^1, X_{t-1}^2, \dots, X_{t-p_2}^2, X_{t-1}^3, \dots, X_{t-p_3}^3, \dots, X_{t-p_3}^m]^T + E_t' \quad (2.7)$$

For more details see Refs. [7, 19].

The interpretation of the second of Granger’s two conditions as a linear regression leads to efficient implementation. It has the drawback, however, that the functional relationships between variables are assumed to be linear and the error term is assumed to be Gaussian. Nonlinear relationships can be included by adding functions of measured variables to the list of predictors [3], but they rely on modeling assumptions and often pit efficiency against modeling precision.

2.2 Shannon Entropy

Entropy can be used to form an alternative interpretation of Granger’s second condition, which is that “the cause and the effect should share information that is not contained in any other variables.” Linear regression can be thought of as a prediction of the mean of a variable, which gives partial information about the variable related to its central tendency. The Shannon or differential entropy of a random variable is a description of the variable that relates to the distribution as a whole. Unlike methods based on linear regression, information-theoretic measures of dependence do not require model specific assumptions about the functional dependence between variables or distributional assumptions. Shannon entropy has a natural interpretation as the information associated with a random variable, and therefore leads to a literal interpretation of Granger’s second condition.

2.2.1 Some intuition about information

Shannon Entropy is a measure of how much “information,” on average, is gained by measuring a random variable. Some simple examples illustrate the meaning of the information content of a random variable. Let us suppose that a math professor teaches a large calculus class, and that the chairman of the department approaches the pro-

fessor and asks if a certain student has been attending lectures. The professor replies that it would be hard to know in a class of that size, but the chairman immediately notes that this student's height is 7'6" (which would be extraordinarily tall at most universities). Such a statement would give the professor a lot more information about who the student is. On the other hand, if the professor had said that the student is 5'8" (roughly average height), the professor would not have learned much. Certain measurements of the variable "height" convey more information than others.

The Shannon entropy of a random variable measures *on average* how much information is learned, where the average is taken over the entire population, or equivalently, over all possible outcomes of the measurement of the random variable. In the case of the height variable, a lot is learned by knowing a student is 7'6", but if most students are roughly average height, then most of the times that a student's height is revealed, not much is learned. Other variables might convey more information on average. For instance, if the students were divided equally among 10 majors, then asking the chairman about the student's major would on average provide a lot of information because the answer would always rule out 90% of the class.

On the other hand, if majoring in math was a prerequisite for taking the class, the answer would not convey any new information. Therefore, it seems that the amount of information that is learned on average is a function of the probability distribution of the variable. In fact it may seem that information is just another way of describing variance, which is the building block for the measures of dependence described in Sec. 2.1.2. If, for example, the heights of students in the class have a large variance, then on average, height could provide a lot of information, but if all of the students are exactly 5'8", so that the variance is 0, no information is gained by asking about height. But there is a big difference. If there were only two outcomes and half of the students were exactly 7'8", and the other half 3'8", then learning the height

would provide much less information than if the heights were uniformly distributed across every value between 3'8" and 7'8". On the other hand, the variance of the first distribution with two outcomes is far greater than the variance of the second, uniform distribution. Instead of being based on variance, entropy is related to the sizes of the partitions created by the variable.

2.2.2 Shannon's treatment of entropy

A mathematical description of the entropy of a random variable was introduced by Shannon in his seminal 1948 paper, "A mathematical theory of communication," in the Bell System Technical Journal [110]. Shannon proposes a set of intuitions which such a measure should possess and demonstrates that these intuitions uniquely characterize, up to a multiplicative constant the entropy of a discrete random variable, meaning a variable with a discrete outcome space.

Theorem 2.2.1 (Shannon's uniqueness theorem). *Suppose that for any random variable X taking values in a finite set $\{x_1, \dots, x_n\}$, there is a unique real number $H(X)$ such that $H(X)$ satisfies the following conditions*

1. *$H(X)$ depends only on the probability distribution of X . Because of this axiom, without loss of generality, $H(X)$ can be written $H(p_1, \dots, p_n)$, where $p_i = \mathbb{P}(X = x_i)$, and x_1, \dots, x_n is an enumeration of the elements in the range of X .*
2. *(Continuity) For a fixed n , among variables X with n possible outcomes with probabilities p_1, \dots, p_n , H is a continuous function of (p_1, \dots, p_n) .*
3. *(Monotonicity) If X_1 has n_1 possible outcomes, each with probability $1/n_1$, and X_2 has n_2 possible outcomes, each with probability $1/n_2$, then $H(X_1) > H(X_2)$ if $n_1 > n_2$.*

$$4. \text{ (Recursivity) } H(p_1, p_2, \dots, p_n) = H(p_1 + p_2, p_3, \dots, p_n) + (p_1 + p_2) H\left(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2}\right).$$

Then there is some positive $\alpha \in \mathbb{R}$ such that

$$H(X) = -\alpha \sum_i p_i \log p_i. \quad (2.8)$$

It is customary to take α to be 1, giving the following definition

Definition 2.2.1 (Shannon Entropy). The Shannon entropy of a discrete random variable, X , taking values in $\{x_1, \dots, x_n\}$ with probability mass function $\mathbb{P}(X = x_i) = p_i$ is

$$H(X) = - \sum_{i=1}^n p_i \log p_i. \quad (2.9)$$

The second condition states that if two probability distributions are close to each other (in any \mathbb{R}^n norm), then their entropies are close to each other. The third condition states that among variables that break the population into equal sized partitions, more partitions means more information. The final condition is called recursivity. Suppose that X can take values in $\{1, 2, 3, \dots, n\}$. Then X could be rewritten in terms of two variables, X_1 and X_2 , where X_1 chooses from the $n - 1$ choice $\{\{1, 2\}, 3, \dots, n\}$, and X_2 chooses between 1, and 2. The fourth statement expresses $H(X)$ as a weighted sum of $H(X_1)$ and $H(X_2)$.

Despite the success of Shannon's paper, it is somewhat dissatisfying to some mathematicians. It is mostly stated in terms of communication theory, even though the results should apply to a broad mathematical context. When one tries to translate the communication theory language about sources, transmitters, lines, channels, codes, and receivers, into the language of probability theory as expressed in terms of Appendix 1, a number of ambiguities and confusions arise. For instance, it is of-

It is unclear from Shannon's exposition whether an object should be interpreted as a random variable or as a sequence of measurements of a random variable, and the difference sometimes affects the interpretation of the text. A number of more purely mathematical treatments of Shannon's ideas have arisen. The earliest attempt was given by the Russian mathematician Khinchin in 1957 [63], who introduces a different set of axioms for Shannon Entropy. Since then many more axiomatizations of Shannon Entropy have been proposed [27]. The different choices of axioms correspond to different characterizations of Shannon entropy.

Similar quantities describing information content can be defined when there is more than one variable. For instance, the joint entropy describes the information gained on average by measuring a pair of variables.

Definition 2.2.2 (Joint Shannon entropy). Given two random variables X and Y taking values in $\mathcal{X} = \{x_1, \dots, x_{n_X}\}$ and $\mathcal{Y} = \{y_1, \dots, y_{n_Y}\}$, the joint entropy of X and Y is the entropy of the variable (X, Y) , that produces pairs of observations, (i, j) , with probability $p_{i,j} = \mathbb{P}(X = x_i, Y = y_j)$. Therefore the joint entropy is

$$H(X, Y) = - \sum p_{i,j} \log p_{i,j}, \quad (2.10)$$

where the sum ranges over $(x, y) \in \mathcal{X} \times \mathcal{Y}$.

The conditional entropy of X given Y can be defined in a similar fashion.

Definition 2.2.3 (Conditional Shannon Entropy). Given two variables X and Y taking values in $\mathcal{X} = \{x_1, \dots, x_{n_X}\}$ and $\mathcal{Y} = \{y_1, \dots, y_{n_Y}\}$, define

$$p_{i|j} = \mathbb{P}(X = x_i | Y = y_j). \quad (2.11)$$

Then the conditional entropy of X given Y is

$$H(X|Y) = - \sum p_{i,j} \log p_{i|j}. \quad (2.12)$$

Conditional entropy can be interpreted as the amount of uncertainty involved in measuring X given that the value of Y has already been revealed. Probabilities of events can be 0, in which case, the convention is made that $0 \log 0 = 0 \log \frac{0}{0} = 0$. These conventions are justified by the continuity axiom characterizing Shannon entropy. Note that for discrete variables $p_{i|j} = \frac{p_{i,j}}{p_j}$ so that

$$H(X|Y) = - \sum p_{i,j} \log \frac{p_{i,j}}{p_j}. \quad (2.13)$$

For reference, Table 2.1 records a list of useful properties of H that follow from the axioms characterizing Shannon entropy. These properties are very intuitive for a measure of uncertainty. For instance, positivity is a statement that a variable cannot have negative uncertainty. In fact, the most certainty occurs when there is a k such that $p_k = 1$, which implies $H(X) = 0$. Invariance means that relabeling, or permuting the symbols will not affect the uncertainty in a measurement. Additivity states that the uncertainty in simultaneously measuring two independent variables (as described in Appendix 1, independence between X and Y is denoted $X \perp\!\!\!\perp Y$) is the same as the sum of uncertainties involved in measuring each variable separately. The Chain Rule, sometimes called Strong Additivity, states that the entropy contained in two variables is equal to the entropy in one of the variables plus the entropy in the other conditioned on knowledge of the first variable. The independence bound states that if two variables are not independent, then together they convey less uncertainty than two independent distributions with the same probability mass functions. The fact that conditioning reduces uncertainty simply states that if there are two variables,

Positivity	$H(X) \geq 0$	*
Invariance	$H(f(X)) = H(X)$ if f is one-to-one	*
Additivity	$X \perp\!\!\!\perp Y \implies H(X, Y) = H(X) + H(Y)$	
Chain Rule	$H(X, Y) = H(X) + H(Y X)$	
Independence Bound	$H(X, Y) \leq H(X) + H(Y)$	
Conditioning Reduces Entropy	$H(X Y) \leq H(X)$	

Table 2.1: Properties of Shannon Entropy. A “*” symbol indicates that the property is not shared by differential entropy (see Sec. 2.3).

which are possibly related, then knowing the value of one can only decrease the uncertainty in the measurement of the other.

2.2.3 Alternative ways of thinking about entropy

The axiomatic approach to defining entropy is useful for establishing intuition. There may be other ways to arrive at the definition of entropy. Alternative frameworks have the potential to lead to new ways of thinking about entropy, and may suggest different generalizations.

The following thread of thought introduces entropy in a way that connects it to a familiar concept from algebra, the geometric mean. There are different ways to assign algebraic structure to probability distributions and probabilities of events. Unlike random variables, probability distributions cannot be added, and even adding probabilities of individual events is of limited significance. On the other hand, multiplication is a natural operation for use with probabilities in many circumstances. For instance, the probabilities of independent events multiply, and in particular, if $\{X_i\}_{i=1}^{\infty}$ is a sequence of independent discrete variables, then

$$\mathbb{P}(X_1 = x_1, X_2 = x_2, \dots, X_d = x_d) = \mathbb{P}(X_1 = x_1)\mathbb{P}(X_2 = x_2) \cdots \mathbb{P}(X_d = x_d). \quad (2.14)$$

Although the average in the definition of Shannon entropy is an average over all

possible outcomes, it could equally be considered an average over many independent samples, because over time the empirical distribution converges to the probability distribution.

A different way of thinking about the information of an event as described in Sec. 2.2.1 is as an amount of “surprise” upon learning the outcome – more “surprising” outcomes imply a greater reduction in uncertainty, and therefore a greater gain in information. Surprise can be defined precisely and shown to have a natural multiplicative structure like probability. It is reasonable that a measure of surprise at an event should be inversely related to the probability of the event. So in particular, the surprise at measuring $X = x_k$ is proportional to $1/p_{x_k}$. One can define variables P_i and S_i such that $P_i(\omega) = \mathbb{P}(X_i = X_i(\omega))$, $S_i(\omega) = 1/P_i(\omega)$, where P_i is recognized as the probability mass function and S_i can be called surprise. Together with Equation (2.14) this definition of surprise states that in the context of making many measurements of the same variable, surprise multiplies:

$$S_{1,\dots,d} = \prod_{i=1}^d S_i. \quad (2.15)$$

Entropy can be thought of as the *average* surprise on measuring a variable. The mathematically appropriate way to average objects that multiply is the geometric mean, which is sometimes written as the n th root of a product of the objects. Another way to write the geometric mean of a variable Y is

$$\text{GM}(Y) = f^{-1} \mathbf{E}[f(Y)] \quad (2.16)$$

where $f(x) = \log(x)$, so that $f^{-1}(x) = e^x$. This way of writing the geometric mean might seem unnecessary at first, but it actually generalizes a wide class of means. For

instance taking f to be the identity yields the arithmetic mean and taking $f(x) = 1/x$ yields the harmonic mean.

Substituting the surprise, S , for Y in Eq. 2.16 yields

$$\text{GM}(S) = \exp \left(\mathbf{E} \left[\log \frac{1}{P} \right] \right) \quad (2.17)$$

$$= \exp (-\mathbf{E} [\log P]) \quad (2.18)$$

$$= \exp \left(-\frac{1}{n} \sum_{k=1}^n p_k \log p_k \right). \quad (2.19)$$

Since we are only interested in the rate at which this quantity scales we define $H(X)$ to be the exponent

$$H(X) = -\frac{1}{n} \sum_{k=1}^n p_k \log p_k, \quad (2.20)$$

which is identical to Def. 2.2.1.

2.3 Extending Shannon entropy to differential entropy

One of the main goals of Shannon's 1948 paper was to create a theory of communication that applied equally well to continuous and discrete signals, as well as signals containing mixtures of continuous and discrete components. This section discusses attempts at extending Shannon's theory for discrete random variables to the more general setting of continuous and mixed variables. For simplicity, in this section a continuous random variable is an \mathbb{R}^d -valued variable with a probability density function (pdf).

The most straightforward approach is to apply the intuitions that define Shannon

entropy in the discrete case. The axioms that uniquely define Shannon entropy for discrete variables do not generalize easily to include continuous variables. For instance, there is no distribution with constant probability density on \mathbb{R} or \mathbb{R}^d , which makes the third of Shannon's axioms, monotonicity, difficult to interpret.

A different approach to the extension is to treat the summation in the definition of Shannon entropy, Def. (2.2.1), as a Riemann summation. As an example, suppose X is an \mathbb{R} -valued random variable taking values in $[0, 1]$. Then one might try to define the entropy of X by considering the probabilities of partition elements $\left[\frac{i-1}{N}, \frac{i}{N}\right]$ as $N \rightarrow \infty$:

$$H(X) = - \lim_{N \rightarrow \infty} \sum_{i=1}^N \mathbb{P} \left(X \in \left[\frac{i-1}{N}, \frac{i}{N} \right] \right) \log \mathbb{P} \left(X \in \left[\frac{i-1}{N}, \frac{i}{N} \right] \right) \quad (2.21)$$

The problem is that this sum is not likely to converge, which can be demonstrated by assuming that X has a probability density function, f_X . Then $\mathbb{P} \left(X \in \left[\frac{i-1}{N}, \frac{i}{N} \right] \right) \approx \frac{1}{N} f_X(x_i)$, where $x_i \in \left[\frac{i-1}{N}, \frac{i}{N} \right]$, so that

$$H(X) = - \lim_{N \rightarrow \infty} \sum_{i=1}^N \frac{f_X(x_i)}{N} \log \frac{f_X(x_i)}{N} \quad (2.22)$$

$$= - \lim_{N \rightarrow \infty} \sum_{i=1}^N \frac{f_X(x_i)}{N} \log f_X(x_i) + \sum_{i=1}^N \frac{f_X(x_i)}{N} \log N, \quad (2.23)$$

and assuming the first term has a finite limit, it can be removed from the limit, yielding

$$H(X) = - \int_0^1 f_X(x) \log f_X(x) dx + \lim_{N \rightarrow \infty} \sum \frac{f_X(x_i)}{N} \log N \quad (2.24)$$

$$\approx - \int_0^1 f_X(x) \log f_X(x) dx + \lim_{N \rightarrow \infty} \log N \quad (2.25)$$

$$= \infty, \quad (2.26)$$

where the integral is interpreted as a Riemann integral. Since this approach yields $H(X) \equiv \infty$ it is not useful as a definition of entropy for continuous random variables.

Another approach, which was taken by Shannon, is to notice that the left hand term in Eq.(2.25) looks a lot like a continuous version of Shannon entropy. He more generally defined entropy for an \mathbb{R} -valued random variable (not necessarily taking values in $[0, 1]$) with a probability density function, f_X as

$$H(X) = - \int_{-\infty}^{\infty} f_X(x) \log(f_X(x)) dx. \quad (2.27)$$

Eq. (2.27) is commonly referred to as differential entropy.

Unfortunately, Equation 2.27 does not satisfy many of the intuitions of Sec. 2.2. For instance, with this definition, $H(X)$ is not always positive, so that it would appear that learning the value of a random value could impart a negative amount of information⁴. An example is given by the variable that is uniform on the interval $[0, a]$, where $a < 1$. By Def. 2.2.1 the entropy is $\log(a)$, which is negative. As another example, if the heights of the students in Sec. 2.2.1 were modeled using a normal distribution, and the normal distribution had a small enough variance, then we might lose information by learning the height of a student.

Another difference from Shannon intuition is that one-to-one transformations of the variables can change their information. For instance, a variable which is uniform on $[0, 1]$ has entropy equal to 0. Multiplying the variable by $a \in (0, 1)$ produces a variable with entropy $\log(a) < 0$. It should be noted, however, that as a consequence of the change of random variables theorem (see Appendix 1), given an invertible linear

⁴Some scientists avoid this problem by focusing on the variable $e^{H(X)}$ which is necessarily positive.

transformation A , Eq. 2.27 satisfies

$$H(AX) = H(X) + \log |A|, \quad (2.28)$$

where $|\cdot|$ indicates determinant. Also, Eq. 2.27 satisfies translation invariance,

$$H(X + c) = H(X). \quad (2.29)$$

But translation seems to be the only operation that is invariant under Eq. (2.27), and for more general transformations it is difficult to find simple formulas like Eq. (2.28).

Another problem with interpreting Eq. (2.27) as an extension of Shannon entropy to continuous random variables is that given an H that is defined for the subspace of discrete variables as well as the space of absolutely continuous variables, it should be apparent how to compute H for variables which have distributions that can be described as mixtures of discrete and continuous parts. The recursivity axiom describes how to compute the entropy of variables which are formed from sampling two variables (Shannon calls this making two choices), so a suitable generalization of the axiom might determine the entropy of the mixture variable. But, consider a distribution such as a variable X with cdf

$$F_X(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{2} + \frac{1}{2}x & x \in [0, 1] \\ 1 & x > 1 \end{cases} \quad (2.30)$$

This variable could be decomposed into two choices in a number of ways. For instance, first one could choose uniformly from the discrete set $\{0, 1\}$, then if the outcome is 1, sample uniformly from the interval $(0, 1)$. Another way would be to sample uniformly

from $[-1, 1]$, and if the answer is less than 0, round up to 0. Although it is not clear what generalization of the recursivity axiom Shannon would have used, it seems that some interpretations could lead to inconsistent derivations of $H(X)$. This might not be surprising after learning that Eq. 2.27 lacks the same interpretation as Shannon entropy as a measure of information or uncertainty in a random variable, but it is interesting that it seems that there might not be any known extension of Shannon Entropy to a functional that gives consistent results on mixed distributions. The derivation of an entropy functional that works in the space of distributions spanned by the discrete and continuous subspaces remains a challenging open question.

Yet another issue is that the definition in Eq. 2.27 only holds for distributions which are absolutely continuous with respect to Lebesgue measure (See Appendix 1 for a definition of absolute continuity and a discussion of the Radon Nikodym theorem). One drawback to using this domain is that it is not a vector space, a fact that is easily verified, for instance by considering that the sum of X and $-X$ would be a discrete variable taking values only at the origin. In fact, the space of absolutely continuous variables does not seem to be closed under any of the algebraic operations it inherits from the range space. Although there are some analytic results regarding the behavior of differential entropy under algebraic operations such as Eq. 2.28, the lack of closure axioms for the algebraic operations makes it very difficult to do any type of analysis to elucidate the general behavior of H with respect to algebraic combinations of variables.

The difficulty in analysis of algebraic combinations of variables is very different from the discrete case. Shannon entropy is subadditive in the sense that

$$H(X + Y) \leq H(X) + H(Y).$$

This subadditivity property is shared by power functions, norms, and concave functions, and is a widely used and powerful property in analysis. For differential entropy the inequality could point in either direction depending on the choice of X and Y . There has been some work on establishing bounds on the differential entropy of sums, but the take-home message is that working with the differential entropy of a sum is very difficult. For instance, Cover and Zhang [26] obtain results using the assumption that the log of the pdfs of the random variables be concave. Madiman and Kontoyiannis [76] show that if X and Y are independent then

$$\frac{1}{2} \leq \frac{H(X+Y) - H(X)}{H(X-Y) - H(X)} \leq 2, \quad (2.31)$$

meaning that the change in $H(X)$ that results from adding Y to X is of the same magnitude, up to a factor of 2, as the change that results from subtracting Y . Entropy power inequalities such as Shannon's

$$e^{2H(X_1+\dots+X_n)} \geq \sum_{j=1}^n e^{2H(X_j)} \quad (2.32)$$

can be used to provide a lower bound on $H(X+Y)$ when X and Y are independent

$$H(\tilde{X} + \tilde{Y}) \leq H(X + Y), \quad (2.33)$$

where \tilde{X} and \tilde{Y} are independent multivariate normally distributed variables chosen such that $H(\tilde{X}) = H(X)$ and $H(\tilde{Y}) = H(Y)$ [31]. Inequality (2.32) relates the squares of the geometric means as defined in Eq. (2.16). A lower bound for independent variables which is simpler to state, but perhaps not very strong is

$$H(X+Y) \geq H(X). \quad (2.34)$$

The point is that even though addition and subtraction are perhaps the simplest algebraic operations used to produce new variables from old (arguably simpler than multiplication in this context, since Lebesgue measure is in some sense uniform with respect to the additive group of the range space), exact analysis is difficult, and even bounding the entropy from above and below is a subject of current research, and may require additional assumptions such as independence.

Another challenge lies in interpreting the meaning of Shannon’s continuity axiom for probability distributions on \mathbb{R}^d . The version of H as defined in Eq.(2.27) is not necessarily continuous. For instance, in a 2016 article in *IEEE Transactions on Information Theory*, Polyanskiy and Wu demonstrate that in the topology generated by Kullback-Leibler distances, differential entropy is not continuous on the space of absolutely continuous distributions [94]. They show that differential entropy is continuous with respect to Wasserstein distance, but only if one restricts to a subset of absolutely continuous distributions defined by a regularity condition. In 2017 in *IEEE Communications Letters*, Ghourchian *et al.* demonstrate that differential entropy can be viewed as continuous with respect to the total variation distance if one restricts to a different subset of absolutely continuous distributions [42].

Despite these problems and challenges with interpreting Eq. (2.27) as an extension of Shannon entropy, Eq. (2.27) has the mathematical form that is expected of an entropy, and is widely used as a measure of entropy of continuous variables. However, because it should not be interpreted as a measure of information or surprise, because it does not satisfy the Shannon axioms, and because it does not extend to variables which have both discrete and continuous components, it is important not to use the same name. A commonly adopted convention is that the term Shannon entropy should only be used in conjunction with discrete random variables, and the quantity (2.27) should be called “differential entropy.” The use of the term differential entropy implies

that the variable under consideration is absolutely continuous with respect to the Lebesgue measure on the range space. Often, the notation h is used for differential entropy to distinguish it from Shannon Entropy. In this thesis the notation H is used for both types of entropy, but the context (discrete or continuous) will indicate which entropy is intended.

2.4 Absolute continuity, Kullback-Leibler divergence, and united framework for Shannon and differential entropies

This section presents an alternative viewpoint on Shannon entropy and differential entropy in which the primary focus is the relationship of absolute continuity between measures and the Kullback-Leibler divergence associated with this relationship. Although there is some folklore suggesting that such a unification might be possible, it is unclear whether the details have been worked out previously. The unification is based on a definition of Kullback-Leibler divergence that is different from the classical definition used in communications theory. The definition in this section focuses on an absolute continuity relationship and the workhorse behind the proofs of the ensuing theorems is the Radon-Nikodym theorem (see Appendix 1).

The first outcome of this approach to differential entropy is a unified approach to Shannon and differential entropy that rescues differential entropy from some of the problems associated with picturing it as extending Shannon entropy as described in Sec. 2.3. Despite the problems and challenges presented in Sec. 2.3, differential entropy has many positive characteristics and uses. For example, Shannon Entropy and differential entropy share the properties in Table 2.1 which are not marked by a

‘*’. More importantly, though, differential entropy plays a key role in a number of important mathematical theorems, such as the maximum entropy characterization of the normal distribution. In addition, differential entropy plays an important role in physics, for instance by generalizing the classical Heisenberg uncertainty principle [8, 12]. It is shown in this section that these positive results about differential entropy follow as a natural outgrowth of the framework of absolute continuity and the Radon-Nikodym theorem. The classical statements and proofs of these theorems can be found in [25], where they are proved separately for discrete and continuous random variables.

The second outcome of this section is the generalization of Shannon and differential entropy. Section 2.4.3 presents this generalization and discusses some applications to topological groups with a Haar measure and to measures supported on a strange attractor.

2.4.1 Shannon Entropy and differential entropy as Kullback-Leibler divergence

The properties that Shannon Entropy shares with differential entropy follow from a unifying description as Kullback-Leibler (KL) divergence. The KL divergence quantifies the dissimilarity between two measures which are related via absolute continuity. It is based on the Radon-Nikodym theorem, which, given two measures $\nu \ll \mu$, guarantees the existence of a Radon-Nikodym derivative, $\frac{d\nu}{d\mu}$ (see Appendix 1 for details).

Definition 2.4.1 (Kullback-Leibler divergence / relative entropy). Let ν and μ be

σ -finite measures⁵ on a measurable space (Ψ, \mathcal{A}) such that

$$\mu \ll \nu. \quad (2.35)$$

Then the Kullback-Leibler divergence of μ with respect to ν is

$$D(\mu \parallel \nu) = \int_{\Psi} \log \frac{d\mu}{d\nu} d\mu, \quad (2.36)$$

where $\frac{d\mu}{d\nu}$ is the Radon-Nikodym derivative of μ with respect to ν .

The quantity $D(\mu \parallel \nu)$ is also called the relative entropy of μ with respect to ν .

Note that asymmetry in the definition of $D(\mu \parallel \nu)$ reflects the asymmetry in the relationship $\mu \ll \nu$ that it quantifies.

The reason for naming the measure space Ψ in Def. 2.4.1 instead of Ω is that the σ -finite measures ν and μ in the definition for KL divergence are often taken to be the push-forward measures (probability distributions) of \mathbb{P} under different choices of random variables where (Ω, \mathbb{P}) is a probability space that models the system being studied, as described in Appendix 1. Therefore, the notation

$$D(X \parallel \nu), \quad (2.37)$$

or

$$D(X \parallel Y), \quad (2.38)$$

will be used to indicate $D(\mu_X \parallel \mu)$ or $D(\mu_X \parallel \mu_Y)$, appropriately, where μ_X and μ_Y are the probability distributions induced by X and Y .

⁵The σ -finite condition (see Radon-Nikodym theorem in Appendix 1) is satisfied automatically by probability measures.

If it is understood that there is a measure that dominates all other measures, then the Radon-Nikodym derivatives with respect to this measure will often be used in place of ν or μ , yielding notation such as

$$D(f_X || f_Y). \quad (2.39)$$

The following Lemma shows that testing whether $X \ll Y$ is relatively easy.

Lemma 2.4.1. *Let X and Y be Ψ -valued random variables with probability distributions μ_X and μ_Y , that are both dominated by a measure ξ . In other words,*

$$\mu_X \ll \xi \text{ and } \mu_Y \ll \xi. \quad (2.40)$$

Then the following two conditions are equivalent:

1. $\mu_X \ll \mu_Y$
2. $\text{supp } \mu_X \subset \text{supp } \mu_Y$.

Proof. The proof is left as an exercise. □

The generality of Def. 2.4.1 permits a unified treatment of Shannon and differential entropy. It can be shown that both Shannon entropy and differential entropy are examples of KL divergences, up to a sign change.

Example 2.4.1 (Shannon entropy is a KL divergence). Assume that X is a random variable taking values in a measure space, $(\mathcal{X}, \mathcal{P}(X), \xi)$ where \mathcal{X} is finite or countably infinite, and ξ is counting measure (that is $\xi(A) = |A|$). Since \mathcal{X} is at most countably infinite, μ is σ -finite.

Let ν be the probability distribution of X on \mathcal{X} . Note that $\xi(A) = 0$ if and only if $A = \emptyset$, so that every measure on \mathcal{X} , including ν , is absolutely continuous. By the

Radon-Nikodym theorem, there is a function $\frac{d\nu}{d\xi} : \mathcal{X} \rightarrow \mathbb{R}$ such that for all $g : \mathcal{X} \rightarrow \mathbb{R}$,

$$\int_{\mathcal{X}} g(x) d\nu(x) = \int g(x) \frac{d\nu}{d\xi}(x) d\xi(x). \quad (2.41)$$

Since ξ is counting measure, the integral is a discrete sum,

$$\int_{\mathcal{X}} g(x) d\nu(x) = \sum_{x \in \mathcal{X}} g(x) \frac{d\nu}{d\xi}(x), \quad (2.42)$$

so that $\frac{d\nu}{d\xi}$ is recognized as a probability mass function,

$$p(x) \equiv \frac{d\nu}{d\xi}(x). \quad (2.43)$$

Using the probability mass function, $p(x)$, Def. 2.4.1 can be simplified:

$$-D(\nu \parallel \xi) = - \int \log \frac{d\nu}{d\xi} d\nu \quad (2.44)$$

$$= - \sum_{x \in \mathcal{X}} p(x) \log p(x) \quad (2.45)$$

$$= H(X), \quad (2.46)$$

where $H(X)$ is the Shannon entropy of X .

Example 2.4.2 (Differential entropy is a KL divergence). Assume that X as an \mathbb{R}^d -valued random variable. Let λ denote Lebesgue measure on \mathbb{R}^d and let ν denote the probability distribution of X on \mathbb{R}^d . If

$$\nu \ll \lambda \quad (2.47)$$

then there is a probability density function of X , $f : \mathbb{R}^d \rightarrow \mathbb{R}$, defined by

$$f = \frac{d\nu}{d\lambda}, \quad (2.48)$$

meaning that for any ν -measurable $g : \mathbb{R}^d \rightarrow \mathbb{R}$,

$$\int_{\mathbb{R}^d} g(x) d\nu(x) = \int_{\mathbb{R}^d} g(x) f(x) dx. \quad (2.49)$$

(λ is σ -finite because a countable collection of unit cubes would cover \mathbb{R}^d .) Therefore, substitution into Def. 2.4.1 yields

$$-D(\nu || \lambda) = - \int f(x) \log f(x) dx \quad (2.50)$$

$$= H(X). \quad (2.51)$$

The unification of Shannon entropy and differential entropy by Kullback Leibler divergence makes it clear why many of the properties listed in Table 2.1 are shared by Shannon entropy and differential entropy. For instance, the chain rule can be seen as a consequence for the chain rule for relative entropy. Since the chain rule implies the additivity property, both of these properties can be seen as following from the chain rule for relative entropy.

In order to state the chain rule a conditional entropy suitable for differential entropy must be defined. Note that in Def. 2.2.3, $H(X|Y)$ is defined to be the expected value over $p_{i|j}$. This motivates the definition of conditional differential entropy.

Definition 2.4.2 (Conditional differential entropy). Let (X, Y) have a distribution, ν , which is absolutely continuous with respect to Lebesgue measure on \mathbb{R}^d . Then the

conditional entropy of Y given X is defined by

$$H(Y|X) = \int_{\mathbb{R}^d} H(\mathbb{P}(Y|X = x)) d\nu(x), \quad (2.52)$$

where the expectation is with respect to the distribution of the variable X .

Theorem 2.4.2 (Chain rule for KL divergence). *Let X and Y be random variables taking values in a σ -finite measure space (Ψ, ξ) such that*

$$\mu_{X,Y} \ll \xi \times \xi, \quad (2.53)$$

where $\mu_{X,Y}$ is the distribution of (X, Y) . Then

$$D(\mu_{X,Y} \parallel \xi \times \xi) = D(\mu_X \parallel \xi) + \mathbf{E}_{\mu_X}[D(\mathbb{P}_{Y|X=x} \parallel \xi)], \quad (2.54)$$

where $\mu_{X,Y}$ and μ_X are the distributions of (X, Y) and X , respectively.

Proof. The absolute continuity of (X, Y) with respect to μ implies that X , and the variables $\mathbb{P}(Y|X = x)$, are absolutely continuous with respect to μ (See Appendix 1). Let $f_{X,Y}$, f_X , and $f_{Y|X=x}$ be the Radon-Nikodym derivatives of these variables with respect to μ . Also, define μ_X and μ_Y to be the probability distributions of X and Y

(so, for example, $f_X = \frac{d\mu_X}{d\mu}$). Then

$$D(\mu_{X,Y} \parallel \xi \times \xi) = \int \log \frac{d\mu_{X,Y}}{d\xi^2} d\mu_{X,Y} \quad (2.55)$$

$$= \log f_{X,Y}(x, y) d\mu_{X,Y}(x, y) \quad (2.56)$$

$$= \iint f_{X,Y}(x, y) \log(f_X(x)f_{Y|X}(y)) d\xi(x)d\xi(y) \quad (2.57)$$

$$= \iint f_{X,Y}(x, y) \log f_X(x) d\xi(x)d\xi(y) \quad (2.58)$$

$$+ \iint f_{Y|X=x}(y)f_X(x) \log f_{Y|X=x}(y) d\xi(x)d\xi(y) \quad (2.59)$$

$$= \int \log f_X(x) \int f_{X,Y}(x, y) d\xi(y) d\xi(x) \quad (2.60)$$

$$+ \int f_X(x) \int f_{Y|X=x}(y) \log f_{Y|X=x}(y) d\xi(y) d\xi(x) \quad (2.61)$$

$$= D(\mu_X \parallel \xi) + \mathbf{E}_{\mu_X}[D(\mu_{Y|X=x} \parallel \xi)], \quad (2.62)$$

where Eq. (2.61) holds by Fubini's Theorem. \square

The chain rule for KL divergence immediately yields the corresponding chain rules for Shannon and differential entropy by letting ξ be counting measure or Lebesgue measure.

Corollary 2.4.2.1 (Chain rule for Shannon and differential entropy). *Let (X, Y) either be a discrete joint variable or let (X, Y) be absolutely continuous with respect to Lebesgue measure on \mathbb{R}^2 . Then*

$$H(X, Y) = H(X) + H(Y|X). \quad (2.63)$$

Note that if X and Y are independent then $H(Y|X) = H(Y)$, which gives the additivity property for both Shannon and differential entropy as an additional corollary.

The proof that conditioning reduces entropy is also due to the fact that Shannon and differential entropies are Kullback-Leibler divergences, but the proof is best left until after the introduction of mutual information in Sec. 2.5.

2.4.2 Applications of differential entropy

As outlined in Sec. 2.3, there are dangers to interpreting differential entropy as a direct extension of Shannon Entropy. On the other hand, differential entropy is used to state some important and beautiful results in mathematics and physics. These results seem to be a result of differential entropy's status as a KL-divergence. A common theme to these results is that in each case differential entropy is used to compare probability distributions. The following theorem is used to demonstrate each of these results. Separate proofs for the discrete and continuous cases can be found in the textbook of Cover and Thomas [25]. The continuous version is based on two dominating relationships, $\mu \ll \xi$ and $\nu \ll \xi$. By focusing on the chain $\mu \ll \nu \ll \xi$ the following proof unites the discrete and continuous case and avoids having to deal with special cases involving division by 0.

Theorem 2.4.3 (Nonnegativity of KL divergence). *Suppose that ν is a probability measure, μ and ξ are σ -finite measures, and*

$$\mu \ll \nu \ll \xi. \tag{2.64}$$

Then

$$D(\mu \parallel \nu) \geq 0 \tag{2.65}$$

with equality if and only if $\mu = \nu$.

Proof. By the chain rule for the Radon-Nikodym derivative,

$$\frac{d\mu}{d\xi} = \frac{d\mu}{d\nu} \frac{d\nu}{d\xi}. \quad (2.66)$$

On the support of ν , $\frac{d\nu}{d\xi} > 0$ ξ -almost everywhere, so that

$$\frac{d\mu}{d\nu} = \frac{d\mu}{d\xi} \bigg/ \frac{d\nu}{d\xi}. \quad (2.67)$$

By Lemma 2.4.1, $\text{supp } \mu \subset \text{supp } \nu$, so that Eq. (2.67) holds on $\text{supp } \mu$. Using this fact,

$$D(\mu \parallel \nu) = \int \log \left(\frac{d\mu}{d\nu} \right) d\mu \quad (2.68)$$

$$= \int_{\text{supp } \mu} \log \left(\frac{d\mu}{d\nu} \right) d\mu \quad (2.69)$$

$$= \int_{\text{supp } \mu} \log \left(\frac{d\mu}{d\xi} \bigg/ \frac{d\nu}{d\xi} \right) d\mu \quad (2.70)$$

$$= \int_{\text{supp } \mu} -\log \left(\frac{d\nu}{d\xi} \bigg/ \frac{d\mu}{d\xi} \right) \frac{d\mu}{d\xi} d\xi. \quad (2.71)$$

Note that $x \mapsto -\log(x)$ is a convex function. Also note that the expected value of the argument of $-\log$ is finite:

$$\mathbf{E}_\mu \left[\frac{d\nu}{d\xi} \bigg/ \frac{d\mu}{d\xi} \right] = \int_{\text{supp } \mu} \left(\frac{d\nu}{d\xi} \bigg/ \frac{d\mu}{d\xi} \right) \frac{d\mu}{d\xi} d\xi \quad (2.72)$$

$$= \int_{\text{supp } \mu} \frac{d\nu}{d\xi} d\xi \quad (2.73)$$

$$= \nu(\text{supp } \mu) \quad (2.74)$$

$$\leq 1, \quad (2.75)$$

because ν is a probability measure. Therefore Jensen's inequality implies that

$$D(\mu \parallel \nu) \geq -\log \int_{\text{supp } \mu} \left(\frac{d\nu}{d\xi} \bigg/ \frac{d\mu}{d\xi} \right) \frac{d\mu}{d\xi} d\xi \quad (2.76)$$

$$= -\log \nu(\text{supp } \mu) \quad (2.77)$$

$$\geq 0. \quad (2.78)$$

If $\nu = \mu$ then $\frac{d\mu}{d\nu} = 1$, and $D(\mu \parallel \nu) = \int \log(1) d\mu = 0$. Conversely, since $-\log$ is strictly convex, an equality in Eq. (2.76) would imply that there exists a constant c such that $\frac{d\nu}{d\xi} = c \frac{d\mu}{d\xi}$ a.e., which, together with the fact that ν and μ are probability distributions, implies that $\nu = \mu$. \square

This theorem leads to maximum entropy characterizations of many important probability distributions.

Definition 2.4.3 (Density function of multivariate normal distribution). An \mathbb{R}^d -valued variable is said to be multivariate normal if there exists a positive definite $d \times d$ matrix Σ and a vector μ such that the pdf of the variable is

$$f(x) = (\det(2\pi\Sigma) \exp((x - \mu)^T \Sigma^{-1}(x - \mu)))^{-1/2}. \quad (2.79)$$

The notation $X \sim \mathcal{N}(\mu, \Sigma)$ means that X has this density function and X is said to be normally distributed with mean μ and covariance matrix Σ .

The following derivation of the entropy of a multivariate normally distributed variable is similar to the derivation found in the textbook by Cover and Thomas [25] but avoids reliance on manipulation of indices.

Lemma 2.4.4 (Entropy of multivariate normally distributed variable). *If X is an*

\mathbb{R}^d -valued variable such that $X \sim \mathcal{N}(\mu, \Sigma)$, then its differential entropy is

$$H(X) = \frac{1}{2} \log(\det(2\pi e \Sigma)), \quad (2.80)$$

Proof.

$$H(X) = -\mathbf{E} \left[\log \left((\det(2\pi \Sigma) \exp((X - \mu)^T \Sigma^{-1} (X - \mu)))^{-1/2} \right) \right] \quad (2.81)$$

$$= \frac{1}{2} (\log(\det(2\pi \Sigma)) + \mathbf{E} [(X - \mu)^T \Sigma^{-1} (X - \mu)]). \quad (2.82)$$

If it can be shown that $\mathbf{E} [(X - \mu)^T \Sigma^{-1} (X - \mu)] = d$ then the proof is complete because $d = \log(e^d)$, which can be absorbed into the other logarithm in (2.82).

But $(X - \mu)^T \Sigma^{-1} (X - \mu)$ is an \mathbb{R}^1 -valued variable (it is the composition of X with a quadratic form), and its expectation may not be immediately apparent. This situation is remedied by converting the expectation into a trace of the expected value of an \mathbb{R}^d -valued variable that is easier to compute.

$$\mathbf{E} [(X - \mu)^T \Sigma^{-1} (X - \mu)] = \text{tr} (\mathbf{E} [(X - \mu)^T \Sigma^{-1} (X - \mu)]) \quad (2.83)$$

$$= \mathbf{E} [\text{tr} ((X - \mu)^T \Sigma^{-1} (X - \mu))] \quad (2.84)$$

$$= \mathbf{E} [\text{tr} (\Sigma^{-1} (X - \mu)(X - \mu)^T)] \quad (2.85)$$

$$= \text{tr} (\mathbf{E} [\Sigma^{-1} (X - \mu)(X - \mu)^T]) \quad (2.86)$$

$$= \text{tr} (\Sigma^{-1} \mathbf{E} [(X - \mu)(X - \mu)^T]) \quad (2.87)$$

$$= \text{tr} (\Sigma^{-1} \Sigma) \quad (2.88)$$

$$= d, \quad (2.89)$$

Steps (2.84), (2.86), and (2.87) follow from the linearity of trace and the expectation operator, and step (2.85) is because trace acts on the product of two matrices in a

commutative fashion. \square

The following is a characterization of the multivariate normal distribution. The proof is similar to the proof in the textbook by Cover and Thomas [25], but makes use of the generalized form of the Kullback-Leibler divergence as stated in this thesis.

Theorem 2.4.5 (Maximum entropy characterization of multivariate normal distribution). *The density function of the multivariate normal distribution is the unique density that maximizes differential entropy among all densities with a specified mean, μ , and variance, Σ . In other words, if X and Y are absolutely continuous with respect to Lebesgue measure, and $X \sim \mathcal{N}(\mu, \Sigma)$, where $\mathbf{E}[Y] = \mu$ and $\mathbf{E}[(Y - \mu)(Y - \mu)^T] = \Sigma$, then*

$$H(Y) \leq H(X) \tag{2.90}$$

with equality if and only if $Y \sim \mathcal{N}(\mu, \Sigma)$.

Proof. If Y and X have the same distribution then $H(Y) = H(X)$. The remainder of the proof considers the case when Y and X do not have the same distribution. The aim is to prove that $H(Y) < H(X)$.

The goal is to apply Theorem 2.4.3 on the nonnegativity of KL divergence. Let π be the distribution of Y , ν be the distribution of X , and λ be Lebesgue measure. It is given that $\pi \ll \lambda$ and $\nu \ll \lambda$. Since $\text{supp} X = \mathbb{R}^d$, it follows that $\text{supp} Y \subset \text{supp} X$ so that

$$\pi \ll \nu \ll \lambda. \tag{2.91}$$

Furthermore, ν is a probability measure, so that the conditions of Theorem 2.4.3 are fulfilled. Let g be the density of Y and f be the density of X . Then, noting that

KL-divergence is 0 only when the distributions are the same,

$$0 < D(\pi || \nu) \quad (2.92)$$

$$= \int_{\mathbb{R}^d} \log \frac{d\pi}{d\nu} d\pi. \quad (2.93)$$

By the chain rule for Radon-Nikodym and because $\frac{d\nu}{d\lambda} > 0$ on \mathbb{R}^d ,

$$\frac{d\pi}{d\nu} = g/f. \quad (2.94)$$

Therefore,

$$0 < \int_{\mathbb{R}^d} g \log \frac{g}{f} dx \quad (2.95)$$

$$0 = -H(Y) - \int_{\mathbb{R}^d} g \log f dx \quad (2.96)$$

$$H(Y) < - \int_{\mathbb{R}^d} g \log f dx. \quad (2.97)$$

So if $-\int_{\mathbb{R}^d} g \log f dx = H(X) = -\int_{\mathbb{R}^d} f \log f dx$ then $H(Y) < H(X)$. It seems that the proof is almost certain to fail, because g could be different from f . On the other hand, the entropy of the normal distribution is a function of Σ . It might be possible that $-\int g \log f dx$ also only depends on the covariance of Y , which is constrained by assumption to be Σ . It is with this hope that we continue:

$$- \int g \log f dx = -\mathbf{E}_\pi[\log f] \quad (2.98)$$

$$= \frac{1}{2} (\log(\det(2\pi\Sigma)) + \mathbf{E}_\pi [(X - \mu)^T \Sigma^{-1} (X - \mu)]) . \quad (2.99)$$

This equation is recognizable as Eq.(2.82), except that the expectation is with respect to Y . But by Eqs. 2.83-2.86, we see that this term is a function of the covariance

matrix, so that it is the same for X and Y . Therefore, $-\int g \log f \, dx = H(X)$, which proves

$$H(Y) < H(X). \quad (2.100)$$

□

Theorem 2.4.3 can be used to characterize a number of important distributions as the unique distributions that maximize differential entropy for a set of constraints and a specific domain. Some examples are the uniform and the exponential distribution. The proofs of the maximum entropy characterizations are omitted, as they are similar to the proof of Theorem 2.4.5, and can be found in the textbook by Cover and Thomas [25].

Definition 2.4.4 (Uniform distribution). Let $[a, b]$ be an interval in \mathbb{R} , and let $\lambda|_{[a,b]}$ be Lebesgue measure restricted to $[a, b]$. A $[a, b]$ -valued variable X is said to be uniformly distributed, written $X \sim \mathcal{U}(a, b)$ if its probability distribution is

$$\nu = \frac{1}{b-a} \lambda|_{[a,b]}, \quad (2.101)$$

which is easily seen to satisfy $\nu \ll \lambda|_{[a,b]}$ and have density

$$f_X(x) = \frac{1}{b-a}. \quad (2.102)$$

Alternatively, one could consider such a variable to be \mathbb{R} -valued, absolutely continuous with respect to λ , with density $\mathcal{X}_{[a,b]}/(b-a)$.

More generally, let Ψ be a topological space with a Regular measure, μ , and let $A \subset \Psi$ have positive measure, $\mu(A) > 0$. Then $\mu|_A \ll \mu$ and $\mu|_A/\mu(A)$ could be called a uniform distribution relative to μ . However, it seems to make more sense to

call a distribution uniform when the topology and the measure does not differ from place to place. This is certainly true when Ψ is a locally compact topological group and μ is a Haar measure, which generalizes \mathbb{R}^d with Lebesgue measure.

Theorem 2.4.6 (Characterization of uniform distribution as maximum entropy distribution). *Suppose $X \sim \mathcal{U}(a, b)$. Then*

$$H(X) = \log(b - a). \quad (2.103)$$

If Y is any other $[a, b]$ -valued variable which is absolutely continuous with respect to Lebesgue measure, then

$$H(Y) \leq H(X), \quad (2.104)$$

with equality if and only if $Y \sim \mathcal{U}(a, b)$.

Definition 2.4.5 (Exponential distribution). A $(0, \infty)$ -valued variable, X , is said to have an exponential distribution with rate parameter $\lambda > 0$, denoted $X \sim \text{Exp}(\lambda)$ if it is absolutely continuous with respect to Lebesgue measure and has pdf

$$f_X(x) = \lambda e^{-\lambda x}. \quad (2.105)$$

Theorem 2.4.7 (Maximum entropy characterization of exponential distribution). *Suppose $X \sim \text{Exp}(\lambda)$. Then $\mathbf{E}[X] = 1/\lambda$ and*

$$H(X) = 1 - \log(\lambda). \quad (2.106)$$

If Y is any other $(0, \infty)$ -valued random variable which is absolutely continuous with

respect to Lebesgue measure and $\mathbf{E}[Y] = 1/\lambda$, then

$$H(Y) \leq H(X), \quad (2.107)$$

with equality if and only if $Y \sim \text{Exp}(\lambda)$.

It is important to note that the normal, uniform, and exponential distributions are defined here by explicitly stating their distributions. It is only then that the actual characterization can be stated, which are then proved by ansatz. This is not good mathematical practice. In general, the characterization should be stated first followed by the derivation of the functional form. In this case deriving the density of the multivariate normal reduces to an optimization, with constraints, over the space of absolutely continuous distributions. It can be checked that the space of absolutely continuous distributions is convex⁶, which gives some hope that such a derivation is possible. Unfortunately such a derivation is outside of the scope of this thesis.

Recently KL divergence has found a number of uses related to dynamical systems. As an example, KL divergence describes the ability of models to make predictions [65]. For instance, by varying the initial conditions, climatological models produce distributions of predictions for a given time t in the future. KL divergence can be used to compare this distribution to other distributions, such as the background historical distribution. A high divergence means that the model is useful because it conveys information beyond what would be known from background conditions. Under certain assumptions, over time the utility decreases to 0.

⁶Assume $\mu \ll \lambda$, $\nu \ll \lambda$, and $\lambda(A) = 0$. Then for $\alpha \in (0, 1)$, $(\alpha\mu + (1 - \alpha)\nu)(A) = \alpha\mu(A) + (1 - \alpha)\nu(A) = 0$

2.4.3 Generalization

The definition of differential entropy in terms of Kullback-Leibler divergence provides a unified definition for discrete variables and variables defined on \mathbb{R}^d absolutely continuous with respect to Lebesgue measure. It is actually much more general than that.

Definition 2.4.6 (Differential entropy (general)). Let ν be a probability measure on a σ -finite measure space (Ψ, ξ) such that

$$\nu \ll \xi. \quad (2.108)$$

Then the differential entropy of ν with respect to ξ is

$$H_\xi(\nu) = -D(\nu || \xi). \quad (2.109)$$

The use of the word differential is justified even in this abstract setting. The Kullback-Leibler divergence is the expectation of the log of a Radon-Nikodym *derivative*. Even though there are examples where the underlying spaces or measures do not appear to be very continuous, the absolute continuity relationship with another measure permits the definition of the Radon-Nikodym derivative, which allows *differential* entropy to be computed.

Example 2.4.3 (Differential entropy on an attractor). Consider a dynamical system,

$$\dot{x} = f(x), \quad (2.110)$$

in \mathbb{R}^d . Assume that this dynamical system has an attractor, \mathcal{A} with an invariant ergodic measure ν . As examples, \mathcal{A} could be a limit cycle, or a strange attractor.

In these examples, ν may not be absolutely continuous with respect to Lebesgue measure, and therefore not have a differential entropy in the traditional sense. But, it is possible to have another measure, ξ , supported on \mathcal{A} such that $\nu \ll \xi$. In the latter case ν would have a differential entropy with respect to ξ defined by $-D(\nu \parallel \xi)$ so that many of the tools of information theory could be used.

It might be possible to construct such a dominating measure, ξ , in a natural way from the dynamical system, Eq. (2.110). Imagine replacing the right hand side by a unit speed velocity, $f(x)/\|f(x)\|$, so that trajectories follow the same attractor, but at a constant speed. This is analogous to an adaptive numerical solver that, under ideal circumstances, would speed up or slow down in exactly the right manner to ensure that the simulated points end up equidistant along trajectories. Such a measure might be thought of as a “uniform” measure on the attractor, and, assuming $\nu \ll \xi$, the quantity $H_\xi(\nu)$ would describe how much dissimilarity from ξ is introduced by allowing the evolution to occur at the natural speed defined by the dynamical system.

Example 2.4.4 (Haar measure). A locally compact group, G , admits a Haar measure, μ , which is uniform under translations by the group operation. An example of a Haar measure is Lebesgue measure on $(\mathbb{R}^d, +)$. But one could also define a Haar measure, $\mu([a, b]) = \log(b/a)$, on the set $(0, \infty)$ with the operation multiplication. Another example is the space of invertible $d \times d$ matrices with the matrix product, or the space of unitary matrices, $U(n)$, again with the matrix product. See, for instance, the discussion section of Madiman and Kontoyiannis (2010) [76].

There are many times when G -valued random variables are useful. For instance, if one wants to “randomly select” a unitary matrix, then they might be thinking about a $(U(n), \cdot)$ -valued random variable whose probability distribution is normalized Haar measure.

There may be another $U(n)$ -valued variable, X , that does not have the same distribution. For instance, the distribution of X might be highly concentrated in a neighborhood of the identity. If X is absolutely continuous with respect to Haar measure, ξ , then the generalized version of Differential Entropy, $H_\xi(X)$, describes how different the two distributions are.

These examples suggest an interpretation of the generalized differential entropy as a quantification of the asymmetry of a measure with respect to the symmetries of another measure. This interpretation extends nicely to Shannon Entropy, in which case the underlying measure is counting measure, which is symmetric with respect to the group of permutations. The uniform measure has the same group of symmetries, so that it seems correct under this interpretation that the Shannon entropy of the uniform measure is 0.

Under this interpretation, maximum entropy distributions (See Theorems 2.4.5, 2.4.6, and 2.4.7) might be described as the distributions that are the closest to having the symmetries of the underlying measure given a set of constraints. The symmetries do not necessarily form a group. For instance, the Lebesgue measure on $(0, \infty)$ has a semigroup of symmetries consisting of right translations. The exponential distribution with a fixed mean has the same semi-group of symmetries (up to a normalization to match the constraint $\int \nu_X = 1$) described by the “memorylessness” property:

$$\mathbb{P}(X > x) = \mathbb{P}(X > x + y \mid X > y). \quad (2.111)$$

It would be interesting to see if the characterization of differential entropy as a quantification of asymmetry can be directly related to properties of groups, semigroups, and other forms of symmetry.

Absolute continuity establishes a preorder on the space of measures on a measur-

able space in which counting measure is a maximal element, and the zero measure is a minimal element. KL divergence and entropy in effect quantify the qualitative relationship of absolute continuity. It would be interesting to see more generally how KL divergence and entropy behave with respect to the preordering. For instance, if

$$\mu \ll \nu \ll \xi \quad (2.112)$$

are σ -finite, then what is the relationship between $H_\nu(\mu)$, $H_\xi(\nu)$, and $H_\xi(\mu)$?

The framework of absolute continuity and Kullback-Leibler divergence does not solve all of the problems described in Sec. 2.3 that arise when one tries to extend Shannon entropy to spaces of continuous and mixed variables. It does offer a new perspective that might be used to view some of these problems in a new light.

2.5 Mutual information

Although KL divergence can be used to quantify how different two distributions are, it does not say anything about the dependence of the random variables that created them. For instance, if $X_i \sim \mathcal{N}(0, 1)$, $i = 1, \dots, 3$, then $D(X_1 \parallel X_2) = D(X_1 \parallel X_3)$, even though it might be that $X_1 \perp\!\!\!\perp X_2$ whereas $X_3 = X_1$. Nothing can be learned about dependence by applying KL divergence directly to the random variables.

KL divergence can be used to define a score for dependence between random variables. The key is that the dependence between the X_i above shows up in their joint distributions. Let μ_i , $i = 1, \dots, 3$ be the distributions of X_i , and $\mu_{i,j}$ the joint distribution of (X_i, X_j) . Then because $X_1 \perp\!\!\!\perp X_2$, $\mu_{1,2} = \mu_1 \times \mu_2$, but $\mu_{1,3}$ would be concentrated along the diagonal $X_1 = X_3$. Thus, a good measure of dependence might measure how far the joint distribution of two variables are from the product

distribution of two variables with the same measure. One such measure is defined by mutual information. It will be demonstrated in Theorem 2.5.2 that X_1 and X_2 are independent when their mutual information is 0. Unfortunately, the mutual information of X_1 and X_3 cannot be defined in this manner, indicating possible limitations of mutual information. See the textbook of Cover and Thomas for a more classical treatment of mutual information [25]. The statements in this thesis are measure-theoretic in nature and aimed at highlighting the important role of absolute continuity.

Definition 2.5.1 (Mutual information). Let X be a Ψ_1 -valued random variable with distribution μ_X , let Y be a Ψ_2 -valued random variable with distribution μ_Y , let $\mu_{X,Y}$ be the distribution of (X, Y) on $\Psi_1 \times \Psi_2$, and let $\mu_X \times \mu_Y$ be the product measure on $\Psi_1 \times \Psi_2$. If $\mu_{X,Y} \ll \mu_X \times \mu_Y$ then the mutual information between X and Y is defined by

$$I(X; Y) = D(\mu_{X,Y} \parallel \mu_X \times \mu_Y). \quad (2.113)$$

The following proposition shows that if the joint distribution is absolutely continuous with respect to a σ -finite measure on the product space then the mutual information exists.

Proposition 2.5.1. *Let $(\Psi_1, \mathcal{F}_1, \xi_1)$ and $(\Psi_2, \mathcal{F}_2, \xi_2)$ be σ -finite measure spaces. Suppose X is a Ψ_1 -valued random variable with distribution μ_X , Y is a Ψ_2 -valued random variable with distribution μ_Y , and $\mu_{X,Y}$ is the distribution of (X, Y) on $\Psi_1 \times \Psi_2$. If*

$$\mu_{X,Y} \ll \xi_1 \times \xi_2, \quad (2.114)$$

where $\xi_1 \times \xi_2$ is the product measure of ξ_1 and ξ_2 on $\Psi_1 \times \Psi_2$, then the following statements hold.

1. $\mu_X \ll \xi_1$,
2. $\mu_Y \ll \xi_2$,
3. $\mu_{X,Y} \ll \mu_X \times \mu_Y$,
4. $\mu_X \times \mu_Y \ll \xi_1 \times \xi_2$.

The third statement implies that $I(X;Y)$ is well-defined.

Proof. 1. Let $A \subset \Psi_1$ be such that $\xi(A) = 0$. Then $\xi_1 \times \xi_2(A \times \Psi_2) = 0$, so that by the assumption of absolute continuity of the joint distribution, $\mu_{X,Y}(A \times \Psi_2) = 0$. But $\mu_X(A) = \mu_{X,Y}(A \times \Psi_2)$, so that $\mu_X \ll \xi_1$ is proven.

2. Same reasoning as 1.
3. This statement follows from statement 4 and Lemma 2.4.1 because $\text{supp}(\mu_{X,Y}) \subset \text{supp}(\mu_X \times \mu_Y)$.
4. It seems logical that the product measure of two absolutely continuous distributions would be absolutely continuous. To check this, let $\xi_1 \times \xi_2(A) = 0$, and, following the notation in Halmos' textbook on measure theory [51], for each $y \in \Psi_2$ define the section

$$A^y = \{x \in \Psi_1 : (x, y) \in A\}. \quad (2.115)$$

Then, by Appendix 1,

$$\xi_1 \times \xi_2(A) = \int_{\Psi_2} \xi_1(A^y) d\xi_2(y) = 0, \quad (2.116)$$

so that $\xi_1(A^y) = 0$ ξ_2 -a.s. Since $\mu_X \ll \xi_1$, this implies $\mu_X(A^y) = 0$ ξ_2 -a.s. Therefore,

$$\mu_X \times \mu_Y(A) = \int_{\Psi_2} \mu_X(A^y) d\mu_Y(y) \quad (2.117)$$

$$= \int_{\Psi_2} \mu_X(A^y) \frac{d\mu_Y}{d\xi_2} d\xi_2(y), \quad (2.118)$$

which is 0 because the integrand is 0 almost surely with respect to the integrating measure. □

An important property of mutual information, as suggested by the motivation, is that the mutual information is 0 exactly when the variables are independent, and positive otherwise.

Theorem 2.5.2. *Let X and Y be random variables on σ -finite measure spaces such that (X, Y) is absolutely continuous on the product measure space. Then*

$$I(X; Y) \geq 0, \quad (2.119)$$

and

$$X \perp\!\!\!\perp Y \iff I(X; Y) = 0. \quad (2.120)$$

Proof. If μ_X, μ_Y and $\mu_{X,Y}$ are the distributions of X , Y , and (X, Y) , where X and Y are (Ψ_1, ξ_1) and (Ψ_2, ξ_2) -valued variables respectively, then by Prop 2.5.1,

$$\mu_{X,Y} \ll \mu_X \times \mu_Y \ll \xi_1 \times \xi_2. \quad (2.121)$$

Therefore, by Theorem 2.4.3

$$D(\mu_{X,Y} \parallel \mu_X \times \mu_Y) \geq 0, \quad (2.122)$$

with equality if and only if $\mu_{X,Y} = \mu_X \times \mu_Y$, which occurs if and only if $X \perp\!\!\!\perp Y$ (see App. 1). \square

It is called *mutual information* because it can be envisioned as the amount of entropy that is shared by the variables. In particular, the venn diagram in Fig 2.1(a) provides a “geometric” analogy for the relationship of mutual information and entropy. If the entropy of X is represented by the left circle, and the entropy of Y is represented by the right circle, then the mutual information of X and Y is the overlap. These types of visualizations should only be taken as analogies, however, because these entropies can take negative values, for instance, when the reference space is (\mathbb{R}, λ) , and λ is Lebesgue measure. The diagrams can be qualitatively useful, however. For instance, Fig 2.1(a) is suggestive of the following decomposition theorem.

Theorem 2.5.3 (Decomposition of mutual information). *Let X and Y be random variables on σ -finite measure spaces such that (X, Y) is absolutely continuous on the product measure space. As shown in Prop. 2.5.1, X and Y are absolutely continuous with respect to the marginal measure spaces, and so $H(X)$ and $H(Y)$ are well-defined. Then*

$$I(X; Y) = H(X) + H(Y) - H(X; Y). \quad (2.123)$$

Proof. Assume X takes values on $(\Psi_1, \mathcal{F}_1, \xi_1)$ and Y takes values on $(\Psi_2, \mathcal{F}_2, \xi_2)$. Let $\mu_X, \mu_Y, \mu_{X,Y}$ be the distributions of X, Y , and (X, Y) . Prop. 2.5.1 shows that

$\mu_X \ll \xi_1$, $\mu_Y \ll \xi_2$, and $\mu_X \times \mu_Y \ll \xi_1 \times \xi_2$.

$$I(X; Y) = \iint \log \frac{d\mu_{X,Y}}{d(\mu_X \times \mu_Y)} d\mu_{X,Y} \quad (2.124)$$

$$= \iint \log \left(\frac{d\mu_{X,Y}}{d\xi_1 \times d\xi_2} / \frac{d(\mu_X \times \mu_Y)}{d\xi_1 \times d\xi_2} \right) d\mu_{X,Y} \quad (2.125)$$

$$= - \iint \log \left(\frac{d(\mu_X \times \mu_Y)}{d\xi_1 \times d\xi_2} \right) d\mu_{X,Y} + D(\mu_{X,Y} \parallel \xi). \quad (2.126)$$

The second term is $-H(X, Y)$. The argument of the logarithm in the first term can be written

$$\frac{d(\mu_X \times \mu_Y)}{d\xi_1 \times d\xi_2} = \frac{d\mu_X}{d\xi_1} \frac{d\mu_Y}{d\xi_2}. \quad (2.127)$$

See Appendix 1 for a discussion of this fact. Therefore, by applying Fubini's theorem, the first term can be simplified,

$$- \iint \log \left(\frac{d(\mu_X \times \mu_Y)}{d\xi_1 \times d\xi_2} \right) d\mu_{X,Y} = - \int_{\Psi_2} \int_{\Psi_1} \log \left(\frac{d\mu_X}{d\xi_1}(x) \frac{d\mu_Y}{d\xi_2}(y) \right) d\mu_{X,Y}(x, y) \quad (2.128)$$

$$= - \int \log \left(\frac{d\mu_X}{d\xi_1}(x) \right) \int \frac{d\mu_{X,Y}}{d\xi_1 \times d\xi_2}(x, y) d\xi_2(y) d\xi_1(x) \quad (2.129)$$

$$- \int \log \left(\frac{d\mu_Y}{d\xi_2}(y) \right) \int \frac{d\mu_{X,Y}}{d\xi_1 \times d\xi_2}(x, y) d\xi_1(x) d\xi_2(y) \quad (2.130)$$

$$= - \int \log \left(\frac{d\mu_X}{d\xi_1}(x) \right) \frac{d\mu_X}{d\xi_1}(x) d\xi_1(x) \quad (2.131)$$

$$- \int \log \left(\frac{d\mu_Y}{d\xi_2}(y) \right) \frac{d\mu_Y}{d\xi_2}(y) d\xi_2(y) \quad (2.132)$$

$$= H(X) + H(Y). \quad (2.133)$$

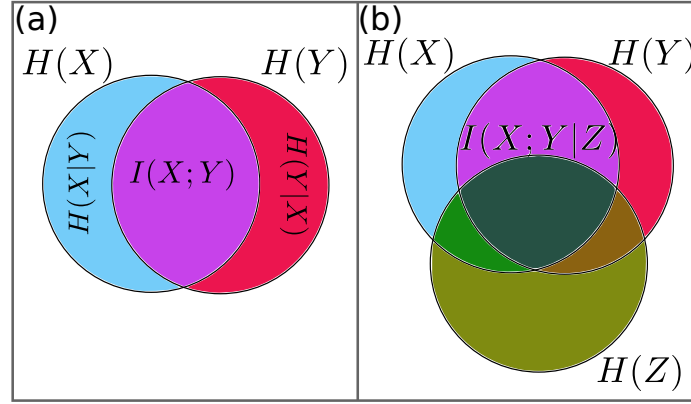


Figure 2.1: Entropy and mutual information as visualized by Venn diagrams. The Venn diagram for entropies should be interpreted with the caution that certain “areas” could be negative.

□

If the variables are discrete, then the mutual information of a variable with itself is simply its Shannon information.

Theorem 2.5.4. *If X is a discrete random variable then*

$$I(X; X) = H(X). \quad (2.134)$$

Proof. Note that the hypothesis of Prop 2.5.1 is trivially satisfied because if ξ is counting measure on Ψ (where X is a Ψ -valued random variable), then $\xi \times \xi$ is counting measure on $\Psi \times \Psi$, and all measures, including $\mu_{X,X}$, are absolutely continuous with respect to counting measure. Therefore $I(X; X)$ is well defined.

Let p_X and $p_{X,X}$ be the Radon-Nikodym derivatives (i.e. the probability mass functions) for X and (X, X) respectively. The theorem follows easily from plugging

in the formula

$$p_{X,X}(x, y) = \begin{cases} p_X(x) & x = y \\ 0 & x \neq y \end{cases} \quad (2.135)$$

to find that $H(X, X) = H(X)$, so that by the decomposition theorem

$$I(X; X) = H(X) + H(X) - H(X, X) \quad (2.136)$$

$$= H(X). \quad (2.137)$$

□

Theorem 2.5.4 does not necessarily hold when the σ -finite measures is not counting measure on a discrete set. For instance, $I(X; X)$ is not even defined when I and H are interpreted as KL-divergences with respect to Lebesgue measure.

Theorem 2.5.5 (Chain rule for mutual information).

$$I(X; Y) = H(X) - H(X|Y). \quad (2.138)$$

Theorem 2.5.6 (Conditioning reduces entropy).

$$H(X|Y) \leq H(X) \quad (2.139)$$

Proof. By the chain rule,

$$H(X) - H(X|Y) = I(X; Y) \quad (2.140)$$

$$\geq 0. \quad (2.141)$$

Therefore $H(X|Y) \leq H(X)$. \square

Conditional mutual information describes the dependence between X and Y that is not already accounted for by a third variable, Z . In Figure 2.1(b), this quantity is depicted as the purple region in the overlap of $H(X)$ and $H(Y)$, but outside of $H(Z)$.

Definition 2.5.2 (Conditional mutual information). Let X , Y , and Z be random variables taking values in σ -finite measure spaces Ψ_X , Ψ_Y , and Ψ_Z . Assume that for each $z \in \Psi_Z$ it holds that the conditional distributions $\mu_{X,Y|Z=z} = \mathbb{P}((X,Y)|Z=z)$, $\mu_{X|Z=z} = \mathbb{P}(X|Z=z)$, and $\mu_{Y|Z=z} = \mathbb{P}(Y|Z=z)$ satisfy the relationship

$$\mu_{X,Y|Z=z} \ll \mu_{X|Z=z} \times \mu_{Y|Z=z}. \quad (2.142)$$

Then the conditional mutual information of X and Y given Z is defined by

$$I(X;Y|Z) = \int D(\mu_{(X,Y)|Z=z} || \mu_{X|Z=z} \times \mu_{Y|Z=z}) d\mu_Z(z), \quad (2.143)$$

2.6 Transfer Entropy

Mutual information quantifies the information shared by two variables, possibly conditioned on the presence of a third variable. If the mutual information is positive, then it can be inferred that the variables are related in some manner, but there is no inherent direction to this relationship, since mutual information is symmetric in its arguments.

If the variables under study are stochastic processes, then time can be used to define the direction of the relationship. Let X and Y be stochastic processes that will be sampled in time, so that the samples can be written (X_1, X_2, \dots, X_n) , and (Y_1, Y_2, \dots, Y_n) , where (X_t, Y_t) is measured before (X_{t+1}, Y_{t+1}) . Given the observation

that the cause should precede the effect (see Sec. 2.1), a relationship of the form $X \rightarrow Y$ would be indicated if in general, X_t and Y_{t+1} shared information.

It is important, though, that the shared information was not already available to process Y . For instance, if the evolution of X is defined by an update rule that depends on Y , but the update rule for Y does not depend on X , then $I(X_t; Y_{t+1})$ might be positive even though knowledge of the state of X_t does not help predict Y_t . In order to exclude information already present in Y , it is necessary to condition on the previous state, or states, of Y . This conditional mutual information is called Transfer Entropy [108].

Definition 2.6.1 (Transfer Entropy). Let X and Y be stochastic processes, and let samples of these processes, indexed by time, be written $\{X_t\}_{t=1}^n$, and $\{Y_t\}_{t=1}^n$. Then the Transfer Entropy from X to Y is defined by

$$T_{X \rightarrow Y} = I(Y_{t+1}; X_t | Y_{t^-}), \quad (2.144)$$

where t^- is meant to indicate that any set of variables of the form $(Y_t, Y_{t-1}, \dots, Y_{t-k})$ could be used as the conditioning set, allowing the flexibility to more aggressively remove any information Y_{t+1} might share with its past.

Furthermore, X_t could be replaced by X_{t^-} since the sharing of information could be delayed. Eq. (2.6.1) is called Transfer Entropy because it is intended to quantify the amount of entropy that is transferred from X to Y because of dependencies created by the evolution of the dynamical process (X_t, Y_t) . It is also referred to as an information flow. A positive $T_{X \rightarrow Y}$ indicates that the process Y does not just depend on Y , but also on X .

Transfer Entropy has been used extensively in scientific applications. For instance, transfer entropy analysis is used in the study of collective animal motion

to infer leader-follower relationships among flying bats from video data [87], and to understand communications between spatially nearby soldier crabs [126]. Transfer Entropy is also used in neuroscience to discover connectivity between different parts of the brain [133]. Other researchers have investigated information transfer between variables descriptive of solar activity to explain the number of sun spots that are present on the surface of the sun at a given time [139]. Refs. [18] and [17] contain a nice exposition on the use of Transfer Entropy to describe the sharing of bits by symbolized versions of dynamical systems during a transition to a synchronized state.

2.7 Causation Entropy

In a networked dynamical systems environment there will generally be more than two variables, which presents problems with interpreting Transfer Entropy as an edge, or a causal relationship. Figure 2.2 depicts two scenarios in which the transfer entropy between X and Y might be positive even though the two nodes are not directly coupled. In Fig. 2.2(a), X only interacts with Y through a set of intermediaries, and in Fig. 2.2(b), a third variable Q drives both X and Y , perhaps with a delay in the driving of Y . Both of these examples set up a statistical relationship between X_t and $Y_{t+\tau}$ for some $\tau > 0$ that cannot be accounted for by conditioning on the past of Y .

The use of Transfer Entropy, therefore, cannot distinguish between direct and indirect edges. It has recently been shown that more generally, any pairwise method, no matter how high its fidelity, will tend to overestimate the number of links in an interaction network, typically resulting in a significant number of false positives that cannot be resolved even with unlimited data [121, 124].

In order to account for these indirect effects, in 2014 Sun and Bollt [121] introduced Causation Entropy (CSE) [20, 64, 73, 121, 123, 124].

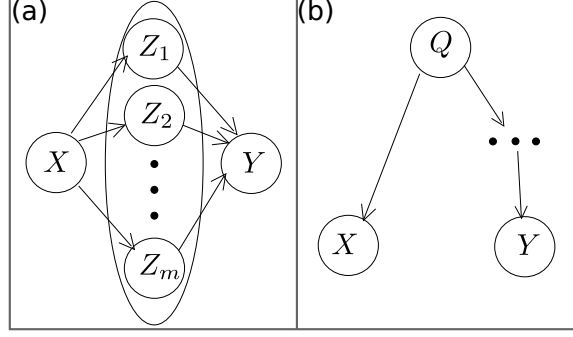


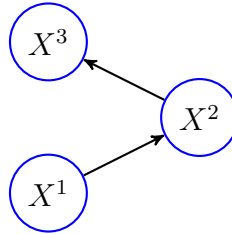
Figure 2.2: Two ways that $I(X^{(t)}, Y^{(t+\tau)} | Y^{(t)})$ could be positive without an edge between X and Y . (a) The variables Z_i (or even a subgraph of $\{Z_i\}_{i=1}^m$) serve as an intermediary between X and Y . (b) The states of X and Y are strongly influenced by a third source, Q . The “...” indicate that there might be other nodes on the path from Q to Y to induce the time lag τ .

Definition 2.7.1 (Causation Entropy). Let $\mathcal{V} = \{X^j : j = 1, \dots, m\}$ be a set of stochastic processes. If $\mathcal{Z} = \{X^{i_1}, \dots, X^{i_k}\}$ is an arbitrary subset of the processes in \mathcal{V} then the Causation Entropy from X^i to X^j conditioned on \mathcal{Z} is

$$C_{X^i \rightarrow X^j | \mathcal{Z}} = I(X_t^i; X_{t+1}^j | X_t^{i_1}, \dots, X_t^{i_k}) \quad (2.145)$$

A positive CSE indicates that X^j depends on X^i in a way that cannot be accounted for by the processes in \mathcal{Z} . Letting $\mathcal{Z} = X^j$ in Def. 2.7.1 recovers Transfer Entropy (Def. 2.6.1) as a special case.

Example 2.7.1. The following example is one of many examples used in the 2014 article by Sun and Bollt to illustrate the differences between Transfer Entropy and CSE. The example concerns a small network of the form



The network is defined by the evolution equations

$$X_{t+1}^1 = f(X_t^1) \quad (2.146)$$

$$X_{t+1}^2 = f(X_t^2) + \frac{\epsilon}{2}g(X_t^1, X_t^2) \quad (2.147)$$

$$X_{t+1}^3 = f(X_t^3) + \frac{\epsilon}{2}g(X_t^2, X_t^3), \quad (2.148)$$

where $f(x) = ax(1 - x)$ is a logistic map, $g(x, y) = f(x) - f(y)$ is a coupling term, and $\epsilon \in [0, 1]$ is a coupling strength. By simulating this system for many time steps the authors are able to estimate the probability density functions associated with each variable and calculate $T_{X^1 \rightarrow X^3}$, and $C_{X^1 \rightarrow X^3 | \{X^2, X^3\}}$. They find that for any coupling strength, $T_{X^1 \rightarrow X^3} > 0$. This makes sense because information is being transferred from X_1 to X_3 indirectly through X_2 . On the other hand, $C_{X^1 \rightarrow X^3 | \{X^2, X^3\}}$ is numerically 0 for all coupling strengths, indicating that all of the information flow from X_1 to X_3 occurred indirectly through the node X^2 .

Given a set of nodes, $\mathcal{V} = \{X^1, \dots, X^m\}$ there is a corresponding set of edges consisting of direct information flows as determined by CSE. In particular,

$$C_{X^i \rightarrow X^j | \mathcal{V} \setminus \{X^i\}} > 0 \quad (2.149)$$

indicates that X^j depends on X^i in a way that cannot be accounted for by any of the other variables, including the past state of X^j . This relationship can be taken as definitive of an edge representing a causal relationship in the CSE sense. It is causal in the sense that it indicates a relationship between two variables that is directed in time and cannot be accounted for by other variables in \mathcal{V} . Furthermore, the information in X^i can be used to improve predictions of X^j . There is, of course, the possibility that non-measured variables outside of \mathcal{V} are influencing X^i and X^j much like the

variables Z_i and Q in Fig. 2.2. If $X^i \rightarrow X^j$ is an edge in the CSE sense then X^i will be called a parent of X^j and X^j will be called a child of X^i .

It should be noted that there is a TE graph consisting of information flows as determined by TE. As example 2.7.1 illustrates, a graph generated in this manner will include the CSE graph as a subgraph, but will also include additional edges which do not correspond to direct interactions between components. TE graphs tend to be much more dense than their corresponding CSE graphs⁷. See Refs. [121, 124] for further discussion.

2.8 The optimal Causation Entropy algorithm

Given a set of nodes, $\mathcal{V} = \{X^1, \dots, X^m\}$, one approach to determining the CSE network of causal edges is to iterate over each pair, (X^i, X^j) , and determine whether $C_{X^i \rightarrow X^j | \mathcal{V} \setminus \{X^i\}} > 0$. This approach would not be practical for even small sizes of \mathcal{V} , partly because there are m^2 ordered pairs of nodes in \mathcal{V} , but also because of the dimension of the distribution of the joint variable that would need to be calculated. If each node was an \mathbb{R}^d -valued random variable, then the joint density involved in the definition of $C_{X^i \rightarrow X^j | \mathcal{V} \setminus \{X^i\}}$ would be defined on $\mathbb{R}^{d(m+1)}$. As described in Ch. 3, the curse of dimensionality can greatly impair the ability to use data to estimate a quantity.

The optimal Causation Entropy algorithm (oCSE) addresses the need for a more practical method of discovering edges [124]. For a given node, X^i , the algorithm begins with an empty set $\mathcal{K} \subset \mathcal{V}$ of parents of X^i . On each step the algorithm finds the j that maximizes $C_{X^j \rightarrow X^i | \mathcal{K}}$, and if that value is greater than 0, adds j to \mathcal{K} , stopping if the maximum is 0⁸. The set \mathcal{K} is now a superset of the parents of X^i [124].

⁷although Example 2.8.1 illustrates a rare occasion in which the TE graph is too sparse

⁸The set \mathcal{K} can be initiated to always contain X^i

The oCSE algorithm then goes through each member of \mathcal{K} and removes an element if $C_{X^j \rightarrow X^i | \mathcal{K} \setminus \{X^j\}} = 0$ because this would indicate that the dependence between X^j and X^i is already accounted for in the other variables in \mathcal{K} . The remaining elements of \mathcal{K} are the causal parents of X^i [124].

In terms of the number of Causation Entropies that need to be computed, the oCSE algorithm does not exceed the brute force approach by much. It is clear that at least m^2 calculations must be performed because for each of the m nodes the first step is to compute all m values $C_{X^j \rightarrow X^i | \emptyset}$. As a worst case scenario, there may be a node for which all other $m - 1$ nodes must be added to \mathcal{K} . Therefore, if $f(m)$ is the number of CSE computations performed, then $m^2 \leq f(m) \leq m^3$.

Whether $f(m)$ is closer to m^2 or m^3 depends on the graph. For instance, in a graph in which a node connects to all other nodes, $f(m) = m^3$. But in a regular graph, or a graph in which the maximum in and out degrees are $k \ll m$, the search is likely to find the causal parents in not much more than km^2 steps. The reason is that indirect influences on nodes are set up through direct influences, so that the number of possible ways to set up indirect influences is limited by the maximum degree of the target node and its parents. If one considers the class of all CSE graphs with maximum in and out degree k , then as $m \rightarrow \infty$ the limit on the ability to create indirect influences does not seem to vary with m , so that $f(m) = \mathcal{O}(m^2)$.

In a more general scenario, one might define a family of graphs by a sparsity constraint. Then there might be a bound on the number of indirect influences whose CSE toward a given node is larger than the CSE of the direct nodes as a function of m . Or, it might be that there is an average bound given some weighting of graphs in the family of size m . In either case, the number of steps might be written as $K(m)m^2$, where $1 \leq K(m) \leq m$. In certain examples, such as discrete time linear stochastic processes with Gaussian distributions [124], $K(m)$ does not seem to depend on the

size of the graph.

These considerations suggest that the number of steps might not be much larger than that of a brute force search. The greatest computational difference between oCSE and the brute force search, however, is the reduced dimensions of the conditioning sets. As Chapter 3 demonstrates, the dimension of the underlying distribution greatly impacts the precision of estimates of quantities associated with the distributions.

An important property of oCSE is that for a broad class of discrete time stochastic processes oCSE exactly recovers the CSE network, and the CSE network is equivalent to the structural network defined by Eq. (2.4) [124]. This class of stochastic processes includes those that have stationary distribution (Def. 2.1.2) and satisfy a set of Markov conditions. Let $X = (X^1, X^2, \dots, X^m)$ be an ordered set consisting of stochastic processes in which the index set is discrete (for example $T = \mathbb{Z}$, or $T = \mathbb{N}$). The conditions that ensure that the CSE network is the same as the structural network and that the CSE network is recovered by the oCSE algorithm are

- Temporally Markov:

$$\mathbb{P}(X_t \mid X_{t-1} = x_{t-1}, X_{t-2} = x_{t-2}, \dots) = \mathbb{P}(X_t \mid X_{t-1} = x_{t-1}) \quad (2.150)$$

for all choices of $(x_{t-1}, x_{t-2}, \dots)$. Colloquially, temporally Markov means that the value of X at time t depends only on the value at time $t - 1$. This would be in contrast to a system with memory or time delays.

- Spatially Markov: If $\{X^{i_1}, \dots, X^{i_k}\} \subset \mathcal{V}$ consists of the structural parents of node X^i then

$$\mathbb{P}(X_t^i \mid X_{t-1} = x_{t-1}) = \mathbb{P}(X_t^i \mid X_{t-1}^{i_1} = x_{t-1}^{i_1}, \dots, X_{t-1}^{i_k} = x_{t-1}^{i_k}). \quad (2.151)$$

for all choices of x_{t-1} .

- Faithfully Markov: Fix X^i . Let $K, L \subset \{1, \dots, m\}$ and let $X^K \equiv \{X^k : k \in K\}$, $X^L \equiv \{X^l : l \in L\}$. Let $N_i \subset \{1, \dots, m\}$ be the indices of the causal parents of X^i . If $K \cap N_i \neq L \cap N_i$ then

$$\mathbb{P}(X_t^i \mid X_{t-1}^K = x_{t-1}^K) \neq \mathbb{P}(X_t^i \mid X_{t-1}^L = x_{t-1}^L). \quad (2.152)$$

on a non-measure 0 subset of the state space.

The significance of the faithfully Markov assumption is best illustrated by an example in which the CSE network is equivalent to the structural network, but the oCSE network fails to find the CSE network.

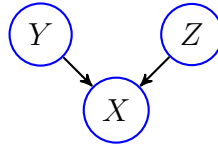
Example 2.8.1 (oCSE does not find CSE graph). This example is introduced by Sun, Taylor and Boltt in 2015 in Ref. [124].

Let $\Psi = \{0, 1\}$ and let $\{Y_t\}_{t \in \mathbb{N}}$ and $\{Z_t\}_{t \in \mathbb{N}}$ be Ψ -valued stochastic processes such that Y_t and Z_t are independent and identically distributed for all t with distribution $\nu(\{0\}) = 1/2$, $\nu(\{1\}) = 1/2$.

Define the stochastic process $\{X_t\}_{t \in \mathbb{Z}_+}$ by

$$X_t = Y_{t-1} \oplus Z_{t-1}, \quad (2.153)$$

Where \oplus denotes the exclusive or operation that returns a 1 if and only if $Y_{t-1} \neq Z_{t-1}$ and a 0 otherwise. The structural graph defined by (2.153) has edges $Y \rightarrow X$ and $Z \rightarrow X$.



It can be checked that $X_t \sim \nu$, and that $X_t \perp\!\!\!\perp Y_{t-1}$ and $X_t \perp\!\!\!\perp Z_{t-1}$. Since the variables Y_{t-1} , Z_{t-1} , and X_t are pairwise independent, and because TE only detects pairwise relationships, the TE graph has no edges, and is therefore too sparse⁹.

The CSE graph, on the other hand, is identical to the structural graph because CSE is able to look beyond purely dyadic relationships. By doing the proper conditioning, Causation Entropy is able to detect that there are interactions which are dependent on other variables. Put in another way, although the variables are pairwise independent, they are not jointly independent, and Causation Entropy detects this dependence by conditioning on the presence of the other variables. It is easy to check from the definition of CSE that

- $C_{Y \rightarrow X|X,Z} = \log(2) > 0$
- $C_{Z \rightarrow X|X,Y} = \log(2) > 0$

and that all other Causation Entropies are 0.

The oCSE algorithm, however, does not find the CSE graph because on the first round all of the transfer entropies are 0 so no extra nodes are added to the conditioning set. The reason is that the faithfully Markov property does not hold because

$$\mathbb{P}(X_t \mid Y_{t-1} = c_1) = \nu = \mathbb{P}(X_t \mid Z_{t-1} = c_2) \quad (2.154)$$

for any $c_1, c_2 \in \Psi$. If this symmetry is broken by setting ν_Y to be slightly different from ν_Z then at least one of the TE's to X will be positive, so that one node will be added, and then, since the CSE conditioning on both nodes is positive, the other node will be added in the second step. Neither node would be removed from the set of parents during the removal phase since both Causation Entropies are positive. It

⁹As described in Sec. 2.7, in practice the TE graphs are typically too dense

is likely that the examples which violate the faithfully Markov property are in some sense non-generic.

Another important result that is proved in Ref. [124] using the same three Markov conditions is the optimal Causation Entropy principle that states that the set of causal parents of a node X^i is the minimal set, $\mathcal{K} \subset \mathcal{V}$, of nodes that maximize the Causation Entropy $C_{\mathcal{K} \rightarrow X^i}$.

2.9 Interpretation of the value of CSE as a real number

Although much has been said in this thesis about the interpretation of positive Causation Entropy conditioned on the remaining nodes as definitive of a CSE edge, not much has been said about the interpretation and use of the numerical value of $C_{Y \rightarrow X|Z}$. The motivation for studying TE and CSE is the notion of an “information flow.” The term “information flow” is used colloquially in the literature, and it might be tempting to take the analogy further and say that the information flows have the properties of physical flows like liquid running through pipes connecting reservoirs. In this analogy the values of TE or CSE might be used to define weights on the edges describing how much “information” is flowing through the edge.

One problem with this interpretation is that differential entropy does not have the same meaning as a measure of “information,” as Shannon entropy (see Sec. 2.3). As described in Sec. 2.4 differential entropy can be thought of as a “divergence” from a reference measure. But to think of a portion of a “divergence” as “flowing” from one node to another is a bit abstract, and should not be expected to behave as a physical flow. Even in the discrete case, the characterization of Shannon entropy

as “information” is equivalent to its characterizations in terms of “uncertainty” or “surprise,” which do not have as much of a physical connotation as “information.” Whether the variables are discrete or continuous, entropy is an abstract quantity which does not necessarily behave like a physical object.

Another problem with such an interpretation is that physical flows are generally defined by a conservation law. The rate at which a substance enters the reservoir should equal the rate at which it leaves plus whatever rate the reservoir might be gaining in volume. In terms of information flows this would imply that the information flow on an edge pointing to a node should be accounted for by either information leaving on outgoing edges or in the information rate of the node. In general, there no such guarantees with CSE or TE, and a quick look at an information diagram 2.1 reveal that there will be areas in the venn diagram that are either not counted (or counted multiple times in the case of TE). In 2016 James, Barnett, and Crutchfield [59] wrote an article for *Physical Review Letters* in which they strongly caution against such an interpretation for a different reason. They state that any definition of an information flow should require that the information be “localizable” to a source, meaning that the information arriving at a target node is “solely attributable” to a source node. They are worried that scientists wanting to apply TE or CSE might misinterpret these quantities as satisfying the localization criterion because of the colloquial use of the term “information flow” in the literature. They use Ex. 2.8.1 and a similar example to illustrate the troubles that can arise from this misinterpretation. In particular, in Ex. 2.8.1, $H(X_t) = \log(2)$, but, $C_{Y \rightarrow X|X,Z} = C_{Z \rightarrow X|X,Y} = \log(2)$. This obviously violates the conservation notion of flow, but they show that the problem arises from a lack of “localizability”. They explain that the Causation Entropies $C_{Y \rightarrow X|X,Z}$ and $C_{Z \rightarrow X|X,Y}$ describe information that is transferred as the result of a “synergy” between Y and Z , that cannot be divided between the sources separately. The authors

say that any formula that is expressible as a conditional mutual information of nodes will fail to be an information flow in the sense of “localizability.”

Instead of discarding the hope of a physical interpretation of information flow, James, Barnett, and Crutchfield suggest, it may be possible to further divide the information shared by the variables into different types of information including information that is unique to each variable, information that is redundant between the variables, and information that arises out of a synergistic effect between variables as described in Refs. [137, 138]. Another possibility that they suggest is to use hypergraphs instead of directed graphs to model the interactions of complex systems. A hypergraph could be written as a triple $(\mathcal{V}, \mathcal{I}, \mathcal{E})$. The set \mathcal{V} has the same interpretation as in the directed graph case – it is the nodes, which, in this thesis, are the stochastic processes. The set \mathcal{I} could be called the set of interactions. The set \mathcal{E} consists of edges (ordered pairs) of the form (v, i) , or (i, v) , where $v \in \mathcal{V}$ and $i \in \mathcal{I}$. Since multiple v can connect to an interaction, the interactions represent polyadic (as opposed to dyadic) relationships. By the variability of its conditioning set, CSE describes aspects of polyadic relationships. It would be interesting to see if CSE would be useful in determining interactions and the information transferred in a polyadic interaction in a complex system as described by a hypergraph.

Chapter 3

Data based inference of causal relationships

An important feature of CSE, which makes it applicable to complex systems research, is that it can be estimated from observational data. As discussed in the introduction, this feature distinguishes it from notions of causality that require intervention [93], as well as methods that require access to the underlying mathematical model that produced the data [71].

The framework for estimation presented here is somewhat unique in that it uses a measure-theoretic presentation to give a unified theoretical background to both parametric and nonparametric statistics. Of particular interest is the organization of much of the theory under the umbrella of the three “strategies” of Sec. 3.1.2. Much of the notation and some of the presentation is taken from Shao’s textbook on mathematical statistics [111]. Some of the notation is borrowed from Lehmann and Casella’s textbook on point estimation [70]. The presentation is not meant to be a comprehensive introduction to estimation, and some important topics have been left out in order to focus more on the nonparametric estimation of differential entropy

and mutual information. The chapter concludes with the presentation of geometric k-nearest neighbors methods given in Warren M Lord, Jie Sun, and Erik M Bollt, “Geometric k-nearest neighbor estimation of entropy and mutual information,” *Chaos: An Interdisciplinary Journal of Nonlinear Science* **28.3** (2018).

3.1 Background

Scientists often need to estimate statistics that describe the distributions of variables that they study. For example, a scientist might need to know the probability of a magnitude 7 earthquake in Los Angeles during a year. Or, they might want to know the mean or median time that a teenager spends on homework during a school week. They might want to know the variance in the circumferences of maple trees in New York. Or they could be interested in the differential entropy of the position of an insect measured periodically in time.

What makes this a challenging problem is that the estimate is made based on a sample consisting of a finite set measurements of the variable. For instance, the scientist interested in the variance of the circumferences of maple trees might make the estimate based on measurements of 100 trees. It seems an understatement to say that values in the sample represent imperfect knowledge about the statistic, or that the sample contains little information. If the underlying distribution of circumferences is absolutely continuous and real valued then the sample can be thought of as representing a measure 0 outcome. It is just one point out of an uncountable number of possible outcomes, and the probability of measuring those exact sample values is 0. In this sense the sample is very unrepresentative of the distribution.

The resolution to this problem is to not think of the sample in terms of the single set of data that might emerge, but to think of the sample as a random variable.

This imparts measurable structure to the potentially uncountable sample space. This measurable structure can be used to define what it means for an estimator to perform well on average across all possible samples that could occur.

3.1.1 The estimation problem as optimization of risk

Let Y be a Ψ -valued random variable where (Ψ, \mathcal{F}_Ψ) is a measurable space. In the above examples, Y might be earthquake size or homework time, for instance. Denote the distribution of Y by ν .

Definition 3.1.1 (Sample). Given a Ψ -valued random variable Y with distribution ν , and a sample size, $n \in \mathbb{Z}_+$, a sample of size n is a Ψ^n -valued random variable,

$$X = (X_1, X_2, \dots, X_n), \quad (3.1)$$

where the distribution of X is ν^n .

The quantity that is being estimated will be denoted $g(\nu)$, or sometimes $g(Y)$, and is called the estimand. For simplicity we will assume here that $g(\nu)$ is real-valued, however, g could take values in a more general measure space. The estimand is defined in terms of a function, g , because the same statistic makes sense for different distributions, and therefore g can be thought of as assigning a value to each distribution. Because most statistics will not make sense for all distributions, ν , and because specific problems call for distributions with specific characteristics, the domain of g will be taken to be a set \mathfrak{F} , which is a subset of the space of distributions. Additional structure will be assigned to \mathfrak{F} as necessary.

Definition 3.1.2 (Estimand). Given a set of distributions, \mathfrak{F} , any real valued function with domain \mathfrak{F} is an estimand on \mathfrak{F} .

Example 3.1.1 (Some typical estimands). If \mathfrak{F} consists of random variables with finite second moments then the variance,

$$g(Y) = \mathbf{E} \left[(Y - \mathbf{E}[Y])^2 \right], \quad (3.2)$$

is an estimand. If \mathfrak{F} is a subset of absolutely continuous variables, the differential entropy,

$$g(\nu) = H(\nu) \quad (3.3)$$

could be an estimand. Of most importance to the application of CSE is the case when Y is assumed to be composed of three random variables whose joint distribution is absolutely continuous,

$$Y = (Y^1, Y^2, Y^3). \quad (3.4)$$

In this case \mathfrak{F} would be a subset of a space of absolutely continuous distributions on a joint space. Then Causation Entropy,

$$g(Y) = C_{Y^1 \rightarrow Y^2 | Y^3}, \quad (3.5)$$

is an estimand.

An estimator of $g(\nu)$ is a measurable function, T , that assigns real numbers to samples.

Definition 3.1.3 (Estimator and estimate). Let (Ψ, \mathcal{F}_Ψ) be a measurable space.

Then T is an estimator if T is a measurable real-valued function on Ψ^n :

$$T : \Psi^n \rightarrow \mathbb{R}. \quad (3.6)$$

If X is a sample, as in Def. 3.1.1, then $T(X)$ is an \mathbb{R} -valued random variable called an estimate. If T is an estimator for a particular estimand g , then T will often be written \hat{g} in order to clarify which estimator the estimand is intended to estimate. Furthermore, if $X = (X_1, X_2, \dots, X_n)$ is a sample for Y , which is distributed as ν , then the estimate will often be written $\hat{g}(Y)$, $\hat{g}(\nu)$, $\widehat{g(Y)}$, or $\widehat{g(\nu)}$ for convenience.

Note that the definition of an estimator does not depend on g and ν . This means that most estimators of $g(\nu)$ are not “useful.” The theory of estimation concerns the problem of finding ways to measure the performance of estimators and designing estimators that perform well with respect to these measures.

The performance of a specific estimate can be measured by a loss function, $L(\nu, T(X))$, where L depends on g , that quantifies the consequences of making a wrong estimate.

Definition 3.1.4 (Loss function). A nonnegative function $L : \mathfrak{F} \times \mathbb{R} \rightarrow \mathbb{R}$ for which $L(\nu, \cdot) : \mathbb{R} \rightarrow \mathbb{R}$ is measurable for any fixed ν is called a loss function for g if for all $\nu \in \mathfrak{F}$ and $d \in \mathbb{R}$,

1. $L(\nu, d) \geq 0$
2. $L(\nu, g(\nu)) = 0$.

Typical examples of loss functions include

$$L(\nu, d) = |g(\nu) - d|^p, \text{ and} \quad (3.7)$$

$$L(\nu, d) = \mathcal{X}_{[c, \infty)}(|g(\nu) - d|), \quad (3.8)$$

where $p, c > 0$ are fixed constants.

Sampling a particular estimate, $T(X)$, for d in $L(\nu, d)$ does not tell the whole story about the estimator, because the data set that the sampling produces might be nonrepresentative of $T(X)$. Since $L(\nu, \cdot)$ is measurable and real-valued, it is possible to find the average loss for a specific ν . The average loss over all possible samples is called the risk of the estimator.

Definition 3.1.5 (The risk of an estimator given a loss function). The risk of using an estimator T when the distribution is ν is defined by

$$R(\nu, T) = \mathbf{E}[L(\nu, T(X))], \quad (3.9)$$

if the expectation exists. This expectation should be interpreted

$$R(\nu, T) = \int_{\Psi^n} L(\nu, T(x)) d\nu(x). \quad (3.10)$$

It should be noted that $R(\nu, T)$ may not exist because the integral may not be finite. This problem can sometimes be addressed by changing \mathfrak{F} or restricting the class of estimators.

If $\nu \neq \nu'$ then generally the risk of using T to estimate $g(\nu)$ will be different from the risk of using T to estimate $g(\nu')$. Unless g happens to be a constant function, there will be no estimator T that gives the lowest risk for every ν^1 . Thus, the goal of developing estimators that uniformly minimize risk over \mathfrak{F} is fruitless. There are three general strategies: constraining \mathfrak{F} , weakening the requirement that T minimize risk for all ν , and restricting consideration to a set of estimators with desirable properties. Usually, more than one of these strategies are combined in order to find a suitable

¹For instance, $T_0 \equiv g(\nu_0)$ has 0 loss by Def. 3.1.4, and therefore minimizes risk for ν_0 , but would have a higher risk for other $\nu \in \mathfrak{F}$.

estimator.

3.1.2 Strategies

Strategy 1: Constrain \mathfrak{F}

Every ν in \mathfrak{F} can be thought of as a constraint, because $R(\nu, T)$ should be smaller than $R(\nu, T')$ for any other estimator. Thus, decreasing the size of \mathfrak{F} makes the problem more tractable. As an example, one may only care that the T be a minimizer for ν that are absolutely continuous with respect to a measure, ξ . The space of absolutely continuous measures with respect to ξ , is still likely to be large since it can be identified with a set of positive $L^1(\xi)$ functions that integrate to 1.

Example 3.1.2 (Parametric families). Historically much of the theory of estimation has focused on parametric families of distributions. In this case, \mathfrak{F} is indexed by a parameter $\theta = [\theta_1, \dots, \theta_m]^T$ from a collection Θ , and can be written

$$\mathfrak{F} = \{\nu_\theta : \theta \in \Theta\}. \quad (3.11)$$

This \mathfrak{F} seems much smaller than the space of absolutely continuous distributions, since it is in some sense locally finite dimensional. In writing expressions involving ν when \mathfrak{F} is a parametrized family, it is traditional to replace ν by the parameter θ . For instance, the estimand is written $g(\theta)$ and the loss function is written $L(\theta, T(X))$.

Example 3.1.3 (Location and Location-scale families). The measurement space Ψ sometimes has symmetries that can be reflected in, or even give rise to, a family of distributions on Ψ . In particular, if G is a group of automorphisms of Ψ , then G acts

on the space of measures of Ψ by

$$g\nu(A) = \nu(g^{-1}A), \quad (3.12)$$

where the inverse of g is used so that if $Y \sim \nu$ then $g\nu$ is the distribution of gY .

If \mathfrak{F} is closed under the action of g , then \mathfrak{F} is called a group family.

The notion of a group family generalizes location, scale, and location-scale families when $\Psi = \mathbb{R}^d$. Some parametric examples are families of normal distributions, double exponential distributions, Cauchy distributions, logistic distributions, exponential distributions, and the family of uniform distributions of the form $\mathcal{U}(a - \frac{b}{2}, a + \frac{b}{2})$ [70].

Strategy 2: Weaken condition that T is minimizer for all $\nu \in \mathfrak{F}$

Unless g is a constant or \mathfrak{F} is extremely small, there will likely not be a T that minimizes $R(\nu, T)$ for all $\nu \in \mathfrak{F}$. There are many ways to weaken this condition in order to obtain a more useful description of the performance of an acceptable T .

Example 3.1.4 (Minimax). One approach is to try to minimize the risk of the worst-case scenario. This is the approach taken by minimax methods, which try to find a T that minimizes

$$\sup_{\nu \in \mathfrak{F}} R(\nu, T). \quad (3.13)$$

Example 3.1.5 (Bayes risk). Another approach is to put weights on the various parts of \mathfrak{F} according to how important it is to get the estimate correct when ν is in that portion of \mathfrak{F} . A little more formally, the Bayesian approach assigns a σ -algebra

to \mathfrak{F} , $\mathcal{F}_{\mathfrak{F}}$, and a probability measure, π to \mathfrak{F} , and tries to minimize

$$\int_{\mathfrak{F}} R(\nu, T) d\pi(\nu). \quad (3.14)$$

In the parametric case, i.e. when $\mathfrak{F} = \{\nu_{\theta} : \theta \in \Theta\}$, the Bayesian approach is equivalent to thinking of Θ as a measurable space, θ as a Θ -valued random variable, and T as the minimizer of

$$\int_{\Theta} R(\theta, T) d\pi(\theta). \quad (3.15)$$

If there exists a T that minimizes Eq. (3.14), then T is called a Bayes estimator.

In practice, π is often given a prior distribution which has maximal entropy on Θ , and a data set sampled from X is used to update π .

Strategy 3: Constrain the set of allowable estimators

It is possible that the estimator that does the best job of minimizing risk brings along with it undesirable properties that one would not expect of a useful estimator. Therefore, one strategy is to demand up front that T belong to a set \mathfrak{T} , which is a subset of the \mathcal{F}_{Ψ} -measurable real valued functions on Ψ^n .

Example 3.1.6 (Equivariance). Suppose two scientists observe the same phenomenon in different coordinate systems. If the scientists are estimating the same quantity using the same estimator then it might be important that they get the same results, or at least that they be able to translate the results between coordinate systems [57]. As a simple example, one scientist might measure temperature in Fahrenheit, and the other in Kelvins. The sample mean is a desirable estimator of the mean in this regard because it commutes with the change of variables, so that estimates in one coordinate

system can be easily translated into estimates in the other coordinate system. This property goes by the name “frame invariance,” or also by “equivariance.”

A little more precisely, let G be a group of automorphisms of Ψ . Suppose that \mathfrak{F} is invariant under G in the sense that for each $\nu \in \mathfrak{F}$, the measure $h\nu$ defined by $h\nu(A) = \nu(h^{-1}A)$ is also in \mathfrak{F} .

The estimation problem would satisfy an equivariance condition if there was a group \tilde{G} acting on \mathbb{R} , such that for any $h \in G$ there is a $\tilde{h} \in \tilde{G}$ such that

$$L(\nu, d) = L(h\nu, \tilde{h}d). \quad (3.16)$$

In this case an estimator is called equivariant (or simply invariant) if

$$\tilde{h}(T\nu) = T(h\nu). \quad (3.17)$$

Equivariant estimation is typically employed when \mathfrak{F} is a location-scale family (See Ex. 3.1.3), and is typically used when $g(\theta) = \theta$.

Example 3.1.7 (Robustness). The field of robust statistics was introduced by Tukey [129], Huber [58], and Hampel [53] in the 1960’s and 1970’s.

Although the sample, X , is defined to be distributed as ν^n in Def. 3.1.1, in real world applications this is rarely a perfect model, and instead, the distribution of X is close enough to ν^n that there *should* be no problem in treating X as a sample of Y . Perturbations from ν^n are called contaminations of the sample. Examples are often divided into two categories.

dispersed: Dispersed contaminations are spread across each of the components of X .

Examples include measurement error, truncation and rounding due to storage on a digital device, and deviations from model assumptions, in particular, the

choice of \mathfrak{F} .

gross: When n is large, a large perturbation to one component of X might still be a small perturbation to X . Causes of gross contamination include the incorrect transcription of data (such as the misplacement of a decimal point), the corruption of bits during data transmission and storage, power surges, and a graduate student sneezing in a lab while a delicate experiment is in progress.

An estimator T is robust if $T(X)$ only changes a small amount due to small contaminations of the data. To make this statement precise would require establishing a topology on the space of measures on Ψ^n , and considering Gateaux derivatives [53].

Although robustness is generally valued, it is not necessary. For instance, the sample mean,

$$T(x_1, x_2, \dots, x_n) = \frac{1}{n} \sum_{j=1}^n x_j, \quad (3.18)$$

is not robust to gross contamination. The sample median is a robust estimator of central tendency [52].

Example 3.1.8 (Unbiased and efficient). A common approach is to require that the estimator be unbiased:

Definition 3.1.6 (Bias). Given a distribution, ν , the bias of an estimator T of $g(\nu)$ such that $T(X)$ has a finite first moment is

$$\text{Bias}_\nu[T] = \mathbf{E}[T(X)] - g(\nu). \quad (3.19)$$

An estimator is called unbiased if $\text{Bias}_\nu[T] = 0$ for all $\nu \in \mathfrak{F}$.

For example, the sample mean is unbiased because if $g(Y) = \mathbf{E}[Y] = \mu$, then

$$\mathbf{E} \left[\frac{1}{n} \sum_{j=1}^n X_j \right] = \frac{1}{n} \sum_{j=1}^n \mathbf{E}[X_j] \quad (3.20)$$

$$= \mu. \quad (3.21)$$

Another desirable property is that $T(X)$ and $T(X')$ are close together when X and X' are both samples of size n of the same variable. This property can be made more formal in terms of the variance of an estimator.

Definition 3.1.7 (Variance of an estimator). Given a distribution ν and an estimate $T(X)$ that has finite second moment, the variance of $T(X)$ is

$$\text{Var}_\nu[T] = \int_{\Psi^n} (T(x) - \mathbf{E}[T(X)])^2 d\nu^n(x). \quad (3.22)$$

Note that the variance of the variables $T(X)$ and $T(X')$ will be the same if X and X' are both distributed as ν^n . Therefore, the notation $\text{Bias}_\nu[T]$ and $\text{Var}_\nu[T]$, which omits X is valid, although it would also be correct to write $\text{Bias}_\nu[T(X)]$ or $\text{Var}_\nu[T(X)]$ for the same quantities.

Definition 3.1.8 (Efficient estimator). An estimator, T , for $g(\nu)$ is called efficient if

$$\text{Var}_\nu[T(X)] \leq \text{Var}_\nu[T'(X)] \quad (3.23)$$

for all other estimators, T' , and all $\nu \in \mathfrak{F}$.

Both the bias and variance of an estimator are related to a commonly used loss function called mean squared error.

Definition 3.1.9 (Mean Squared Error). When $T(X)$ has finite second moment, the risk function corresponding to the quadratic loss in Eq. (3.7) is called mean squared

error:

$$\text{MSE}_\nu[T] = \mathbf{E}[(g(\nu) - T(X))^2]. \quad (3.24)$$

The identity

$$\text{MSE}_\nu[T] = \text{Var}_\nu[T] + \text{Bias}_\nu[T]^2. \quad (3.25)$$

can be derived from Eq. (3.24).

If $\text{Bias}_\nu[T] = 0$ then $\text{MSE}_\nu[T] = \text{Var}_\nu[T]$. In this case, in order to spread the risk over the largest set, it is sensible to search for an estimator T such that

$$\text{Var}_\nu[T] \leq \text{Var}_\nu[T'] \quad (3.26)$$

for any other unbiased estimator, T' , and all $\nu \in \mathfrak{F}$. Such an estimator is called a uniformly minimum variance unbiased estimator (UMVUE).

Example 3.1.9 (Asymptotic statistics). Although some methods for deriving unbiased or efficient estimators exist, they generally only apply to parametrized families of distributions. Furthermore, even when a family can be parametrized, an unbiased or efficient estimator might have other less desirable properties. A commonly used way to weaken the requirements, but retain some of the benefits is by requiring that the properties only hold asymptotically as $n \rightarrow \infty$.

Definition 3.1.10 (Asymptotically unbiased). A sequence of estimators for $g(\nu)$, $T_n : \Psi^n \rightarrow \mathbb{R}$, where $n \in \mathbb{Z}_+$ such that T_n has finite expectation for all n is called

asymptotically unbiased if

$$\lim_{n \rightarrow \infty} \text{Bias}_\nu[T_n] = 0, \quad (3.27)$$

As an example, if ν has a uniform distribution on $(0, \theta)$, then the estimator $T_n(X_1, \dots, X_n) = \max(X_1, \dots, X_n)$ is asymptotically unbiased. One strategy for creating asymptotically unbiased estimators is to start with an asymptotically biased estimator and either divide by its expected value in the limit as $n \rightarrow \infty$ or subtract the asymptotic bias.

As sample size increases the variance of $T(X)$ should decrease, and ideally vanish as $n \rightarrow \infty$.

Definition 3.1.11 (Asymptotic variance). The asymptotic variance of a sequence of estimators for $g(\nu)$, $T_n : \Psi^n \rightarrow \mathbb{R}$ is defined to be $\lim_{n \rightarrow \infty} \text{Var}_\nu[T_n]$ if the limit is finite.

Another criterion is consistency.

Definition 3.1.12 (Consistent estimator). Let $T_n : \Psi^n \rightarrow \mathbb{R}$, $n \in \mathbb{Z}_+$ be a sequence of estimators of $g(\nu)$ and let $X^{(n)}$ be a sequence of samples of size n of ν . Then $\{T_n\}_{n=1}^\infty$ is consistent if $T_n(X^{(n)})$ converges in probability to $g(\nu)$. In other words, T_n is consistent if

$$\lim_{n \rightarrow \infty} \mathbb{P}(|T_n(X^{(n)}) - g(\nu)| > \epsilon) = 0. \quad (3.28)$$

for all $\epsilon > 0$.

One way to check consistency is to verify that the asymptotic bias and variance vanish as $n \rightarrow \infty$.

Theorem 3.1.1. *A sequence of estimators for $g(\nu)$, $T_n : \Psi^n \rightarrow \mathbb{R}$, $n \in \mathbb{Z}_+$, is consistent if the following two conditions hold.*

1. $\lim_{n \rightarrow \infty} \text{Var}_\nu[T(Y_1, \dots, Y_n)] = 0$
2. $\{T\}_{n=1}^\infty$ is asymptotically unbiased for $g(\nu)$.

Proof. By Markov's Inequality (see Appendix) for any $\epsilon > 0$,

$$\mathbb{P}(|T_n(X^{(n)}) - g(\nu)| \geq \epsilon) \leq \frac{1}{\epsilon^2} \int_{\{x \in \Psi^n: |T_n(X^{(n)}) - g(\nu)| \geq \epsilon\}} |T_n(X^{(n)}) - g(\nu)|^2 d\nu \quad (3.29)$$

$$\leq \frac{1}{\epsilon^2} \text{MSE}_\nu[T_n] \quad (3.30)$$

$$= \frac{1}{\epsilon^2} (\text{Var}_\nu[T_n] + \text{Bias}_\nu[T_n]) . \quad (3.31)$$

Taking limits on both sides proves consistency. \square

3.2 Nonparametric estimation of differential entropy and mutual information

3.2.1 Estimation of differential Entropy

Since CSE is a conditional mutual information (CMI), the estimation of CSE can be reduced to the problem of estimating CMI. Before estimating CMI, it is necessary to be able to estimate mutual information (MI). The estimation of mutual information can be reduced to the estimation of differential entropy by the formula

$$I(X; Y) = H(X) + H(Y) - H(X; Y). \quad (3.32)$$

Therefore, this section begins with a discussion of nonparametric methods for estimating $H(X)$. It turns out that the bias of the estimation of $I(X; Y)$ by plugging estimates of $H(X)$ into Eq. 3.42 can be improved upon greatly by performing the estimations in parallel. This method can be extended to CMI.

In the previous section the sample was named X for convenience. Because this section considers joint variables, it is more convenient to switch notation and use X , Y , and Z for the variables whose entropy or mutual information will be estimated. Subscripts on a variable, such as (X_1, X_2, \dots, X_n) will indicate a sample of size n . When there are more than three variables whose joint entropy or mutual information will be estimated, they will sometimes be denoted $X^{(1)}, X^{(2)}, \dots$.

Since differential entropy is defined for absolutely continuous measures on \mathbb{R}^d , it is natural to make the restriction

$$\mathfrak{F} = \{\nu : \nu \ll \lambda, \text{ and } -\infty < H_\lambda(\nu) < \infty\}, \quad (3.33)$$

where λ is Lebesgue measure for the appropriate dimensional Euclidean space. It should be noted that the invariant distributions associated with purely deterministic dissipative dynamical systems are often singular continuous, and so not in \mathfrak{F} , but in the real world situations that these systems model there are noise sources, so that the resulting distribution might be modeled by an absolutely continuous distribution which is in some sense close to the original singular invariant distribution. In short, it is true that some of the distributions that scientists study may not be purely absolutely continuous, but since differential entropy is only defined for absolutely continuous distributions, it does not make sense to extend \mathfrak{F} .

There are many approaches to building nonparametric estimators of differential entropy. Many are based on algorithms to partition the space based on the locations of data points [28, 119]. Another approach is to estimate the pdf of the distribution, and plug that into $H(X) = -\int f \log f$. In one or two dimensions the estimation can be accomplished by kernel density estimation [32, 79, 100] and numerical integration. This approach becomes unfeasible in greater than 2 dimensions [60]. To avoid numer-

ical integration, it is possible to perform binning by dividing Ψ into an array of equal volume rectangular bins, and using the data to create a histogram that approximates the pdf on each bin. This method also suffers the curse of dimensionality, requires very large bins in order for the count of points in each bin to be large enough to get an accurate estimate of f on the bin, creates disparity between the accuracies of the estimates on each bin, and fluctuates with the width of the bins [38]. More modern methods of partitioning the data space include projection pursuit density estimation [38], which provides geometric information about the data, and adaptive partitioning [28, 119], which is designed to be computationally efficient, having a computational complexity of $\mathcal{O}(n \log n)$.

Many estimates of the density, \hat{f} , are combined with a resubstitution estimator, which is a nonparametric estimator that substitutes the sample into a discretized version of the integral [9].

Definition 3.2.1 (Resubstitution). Given an estimate, \hat{f} , of the density of a random variable distributed as ν , and a sample, X , the resubstitution estimator for $H(\nu)$ is defined by

$$T(X) = -\frac{1}{n} \sum_{i=1}^n \log(\hat{f}(X_i)). \quad (3.34)$$

It is called “resubstitution” because the sample is used to obtain estimate \hat{f} , and then the sample is used again, by substitution it for x in $\hat{f}(x)$.

A more geometric approach is based on k -nearest neighbor statistics. The k -nearest neighbors (knn) estimators have received particular attention due to their ease of implementation and efficiency in a multidimensional setting. By the form of the resubstitution estimator in Eq. (3.34), it is only necessary to estimate the pdf at the value of X_i . If the data points are closely spaced near that point then the estimate

should be large, and conversely, if the data points are sparse in that neighborhood then the estimate should be low. One way to express this is that if V_i is a volume containing X_i , where (X_1, \dots, X_n) is a sample of (X, Y) then

$$\hat{f}(X_i) = c \frac{K_i/n}{V_i}, \quad (3.35)$$

where K_i/n is fraction of the remaining data points² contained in the volume and c is a constant that can be chosen to ensure that \hat{f} integrates to 1. If V_i is the volume of a sphere, and $v_d = \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)}$ is the volume of a unit sphere, then

$$V_i = v_d R_i^d, \quad (3.36)$$

where R_i is the radius. If R_i is the distance to the k th nearest neighbor, where $K_i = k$ is a constant, then the density is approximately

$$\hat{f}(X_i) = \frac{k/n}{v_d R_i^d}, \quad (3.37)$$

which, when plugged into the resubstitution estimator yields

$$\hat{H}(X) = \log(n) - \log(k) + \log(v_d) + \frac{d}{n} \sum_{i=1}^n \log R_i. \quad (3.38)$$

This estimator is biased. It can be made asymptotically unbiased by subtracting the

²The quantity K_i/n is technically an estimator of the number of remaining data points. Occasionally authors use $K_i/(n-1)$ since X_i is already accounted for and could not be one of the remaining points.

asymptotic bias, yielding the Kozachenko-Leonenko estimator [67]³

$$\hat{H}_{KL}(X) = \log(n) - \psi(k) + \log(v_d) + \frac{d}{n} \sum_{i=1}^n \log(R_i), \quad (3.39)$$

where

$$\psi(k) = \frac{\Gamma'(k)}{\Gamma(k)} \quad (3.40)$$

$$= H_{k-1} - \gamma, \quad (3.41)$$

is the digamma function, H_k is the k th harmonic number, and $\gamma = -\psi(1)$ is the Euler-Mascheroni constant.

The Kozachenko-Leonenko estimator is asymptotically unbiased and its variance approaches 0 as $n \rightarrow \infty$. Together these properties imply that it is a consistent estimator [112] (See Def. 3.1.12 and Theorem 3.1.1).

The Kozachenko-Leonenko estimator does not appear to be robust (See Ex.3.1.7 for a definition). This is apparent from the $\log(R_i)$ term. As $n \rightarrow \infty$, the variable $R_{(1)} \equiv \min_{i \in \{1, \dots, n\}} R_i$ will approach 0 almost surely⁴. Thus, for a large enough sample size, the estimate involves the log of a number very close to 0. The slope of $\log(x)$ is $1/x$, so that a small change in the minimum R_i could mean a large change in the estimate. A dispersed contamination of the sample would likely change all of the R_i , and in particular the minimum R_i , resulting in a large change to the estimator.

³Since $K_i/(n-1)$ is sometimes used as an estimate in Eq. (3.35), a $\log(n-1)$ will sometimes appear in place of a $\log(n)$ in Eq. (3.39). Kozachenko and Leonenko used $\log(n-1)$ in their 1987 paper [67]. Kraskov, Stögbauer, and Grassberger [68] use $\psi(n)$ which is much closer to $\log(n-1)$ than $\log(n)$. Other sources use $\log(n)$ [41, 112]. The difference is not considered important, since $\log(n)$, $\log(n-1)$, and $\psi(n)$ are asymptotically equivalent and the convergence is very rapid, so that for small sample sizes the difference is overshadowed by the variance of the estimator.

⁴Intuitively, there must be some compact ball, $B(0, s)$, centered at the origin in \mathbb{R}^d that contains a nonzero fraction of the mass of ν , so that for any $\omega \in \Omega$ (See Appendix 1), the Borel-Cantelli lemmas imply that with probability 1 $X_n(\omega)$ will have a subsequence $X_{n_j}(\omega)$ in $B(0, s)$. By the compactness of the ball, $X_{n_j}(\omega)$ has a convergent subsequence, so that $R_{(1)} \rightarrow 0$ almost surely.

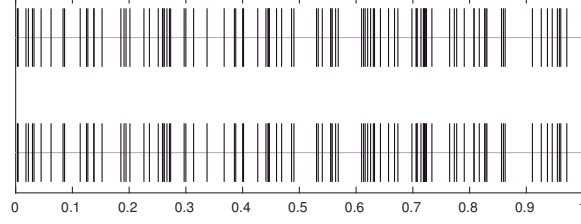


Figure 3.1: The top rug plot shows a sample of size 100 from a uniform distribution on $(0, 1)$. The bottom rug plot shows the same data but with one of the data points perturbed slightly toward its nearest neighbor. The size of the perturbation is on the order of 10^{-5} . The Kozachenko-Leonenko estimates for the entropy are $\hat{H} = -0.02$ (top) and $\hat{H} = -1.03$ (bottom).

Figure 3.1 shows that data sets that appear to the naked eye to be the same can have radically different Kozachenko-Leonenko estimates. In this case the perturbed estimate is below -1 , which would be highly improbable if one sampled a uniform distribution on $(0, 1)$, and yet the data is nearly identical to data sampled from a uniform distribution. This is not proof of non-robustness. In terms of the definition, one would need to start with a uniform distribution on an n -dimensional hypercube, and look for small perturbations, perhaps near the hyperplanes $X_i = X_j$, that would cause the distribution of $\hat{H}(X)$ to change at a rapid rate.

3.2.2 Extension of differential entropy estimators to mutual information

Most strategies for estimating mutual information are based on the formula

$$I(X; Y) = H(X) + H(Y) - H(X, Y). \quad (3.42)$$

Assuming X and Y are \mathbb{R}^{d_X} and \mathbb{R}^{d_Y} -valued random variables, if \widehat{H}_{d_X} , \widehat{H}_{d_Y} and $\widehat{H}_{d_X+d_Y}$ are estimators for $H(X)$, $H(Y)$, and $H(X, Y)$, then an estimator for $I(X; Y)$

is defined by

$$\widehat{I}(X; Y) = \widehat{H}_{d_X}(X) + \widehat{H}_{d_Y}(Y) - \widehat{H}_{d_X+d_Y}(X, Y). \quad (3.43)$$

One method, which is called the copula method [21, 44], makes use of the fact that mutual information is invariant under monotonic transformations of the marginal variables. The distributions of variables in \mathbb{R}^d can be defined by their cumulative density functions (cdf) (see Appendix 1). If F_X is a cdf of a variable X , then the variable $F_X(X)$ is uniformly distributed on $(0, 1)$. Therefore, $H(F_X(X)) = 0$ and, by (3.42),

$$I(X; Y) = H(F_X(X), F_Y(Y)). \quad (3.44)$$

The importance of this way of writing the mutual information is that it moves all of the dependence between X and Y into the joint distribution of a single variable, $(F_X(X), F_Y(Y))$, which has a compact domain. It is not clear, however, that this approach improves the nonparametric estimation of mutual information, because it breaks the estimation up into three estimates. Both F_X and F_Y need to be estimated, in addition to the joint variable $H(F_X(X), F_Y(Y))$. It seems important to avoid using the empirical cdf estimator for the marginal cdfs, defined by

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n \mathcal{X}_{(-\infty, x]}(X_i), \quad (3.45)$$

because then $\hat{F}((X_1, \dots, X_n))$ would consist of an equally spaced lattice of points, which would be extremely unlikely given a uniformly distributed variable, and would therefore have an entropy much less than 0. There are other estimators of cdfs [111], however, once these estimates are performed, the estimation of the joint entropy can

also be challenging (see Sec. 3.2), and there is the possibility of the error from the first estimates feeding forward into the estimate of the joint entropy.

Another approach is to estimate each entropy in (3.42) using the Kozachenko-Leonenko estimator. The resulting estimator for $I(X;Y)$ is called the 3KL estimator [41]. The estimator is asymptotically unbiased since expectation is linear, and it is also easy to check that the variances approach 0 as $n \rightarrow \infty$. However, the variance for a given n might be close to the sum of the variances of each Kozachenko-Leonenko estimator rendering this estimator somewhat impractical. Furthermore, the non-robust terms do not fully cancel.

In 2004 Kraskov, Stögbauer, and Grassberger introduced an estimator based on Eq. 3.43 that cancels much of the bias incurred by making three separate estimates of entropy. The estimator is commonly referred to as the KSG estimator, or just KSG. Instead of calculating the estimates separately, KSG uses information obtained during the calculation of the Kozachenko-Leonenko estimate of $H(X,Y)$ to derive the estimates of the differential entropies of the marginal distributions. In particular, KSG records the radii, R_i , of the spherical volumes in Eq. (3.35) during the estimation of the joint entropy. Instead of using Kozachenko-Leonenko to estimate the marginal entropies, KSG fixes the radius of the volume of the sphere centered at the i th data point to R_i , and treats K_i in Eq. (3.35) as a random variable that depends on how many data point in the marginal space lie inside this sphere. The non-robust terms of the form $\frac{d}{n} \sum_{i=1}^n \log(R_i)$ cancel out. A problem however, is that if Euclidean spheres are used in the joint space, then spheres in the marginal space will tend to be larger than they need to be to hold K_i data points, introducing extra bias. This problem is handled by using a max-norm sphere in the joint space. One of the data points lies on the boundary, ensuring that its projection into the marginal space lies on the boundary of the max-norm sphere projected into the marginal space.

This choice of sphere ensures that in at least one of the marginal spaces R_i is the distance from $\pi(X_i)$ to $\pi(X_{K_i})$, where π is the appropriate projection. The authors offer an alternative version of KSG in which the volume element in the joint space is a hyper-rectangle defined so that the projection of the hyper-rectangle into each marginal space produces a sphere in which the distance from $\pi(X_i)$ to $\pi(X_{K_i})$ is equal to R_i . The use of the hyper-rectangle offers a slight improvement on the bias over the max-norm sphere. The max-norm estimator can be written

$$\hat{I}_{KSG} = \psi(k) + \log(n) - \frac{1}{n} \sum_{i=1}^n (\psi(K_{x,i} + 1) + \psi(K_{y,i} + 1)), \quad (3.46)$$

where k is the fixed number of neighbors used in the joint space and $K_{x,i}$ and $K_{y,i}$ are random variables whose values are number of neighbors of $\pi_x X_i$ and $\pi_y X_i$ in the respective volume elements of radii R_i . Similarly the KSG estimator using hyper-rectangles is

$$\hat{I}_{KSG2} = \psi(k) - \frac{1}{k} + \log(n) - \frac{1}{n} \sum_{i=1}^n (\psi(K_{x,i}) + \psi(K_{y,i})). \quad (3.47)$$

The KSG estimator generalizes to CMI [130], in which case the estimator is

$$\hat{I}(X; Y|Z) = \psi(k) - \frac{1}{n} \sum_{i=1}^n (\psi(K_{xz,i} + 1) + \psi(K_{yz,i} + 1) - \psi(K_{z,i} + 1)), \quad (3.48)$$

where $K_{xz,i}$ and $K_{yz,i}$ are the number of neighbors of the projections of the sample into the marginal spaces spanned by the X and Z coordinate, and the Y and Z coordinates.

Like adaptive partitioning methods [28, 119], KSG's computational complexity using k-d trees [11] is $\mathcal{O}(n \log n)$. However, in practice data sets have limited size, and k-d trees suffer the curse of dimensionality [37]. In order to use K-d trees to

cluster the data in a usable and efficient manner it should hold that $2^d \ll n$. If the data is high dimensional and the sample size is limited then it is of no relevance that the computation time scales asymptotically as $\mathcal{O}(n \log n)$. When data size is limited and dimension is high it is more efficient to use an exhaustive search. In this case, the computation time as n is varied near a maximum reasonable data size scales much more like n^2 . The applications considered in this thesis generally fall into this later category of high dimension and limited sample size.

3.3 A class of knn estimators that adapts to local geometry

This section introduces a new class of nonparametric estimators for differential entropy and mutual information recently introduced in Ref. [72].

3.3.1 Introduction

Under some smoothness conditions on the distributions of X and Y , the KSG estimator is a consistent estimator [41]. However, for finite n , the estimator can be very biased. In fact, as this section demonstrates, given a sequence of joint distributions which are absolutely continuous with respect to Lebesgue measure, it is possible that the bias of the KSG estimator increases unboundedly. In terms of a risk analysis, given a reasonable standard loss function, the minimax of the risk over the space of absolutely continuous distributions with finite entropy, \mathfrak{F} , is infinite. It is possible that by weighting \mathfrak{F} appropriately, that is, treating the risk as a Bayes risk by assigning a measure to \mathfrak{F} , one could make the expectation of the risk finite (See Ex. 3.1.5). This section demonstrates that such a measure would have to give a small weight to

areas of \mathfrak{F} in which distributions of variables that are of great importance in non-linear dynamics and its applications are expected to lie. The findings apply to k nn estimation in general.

One of the most striking geometric features of attractors common to a wide class of dynamical systems, including dissipative systems and dynamical systems with competing time scales, is stretching and compression in transverse directions [88]. More precisely, the local geometry is characterized by both positive Lyapunov exponents corresponding to directions in which nearby points are separated over time and negative Lyapunov exponents corresponding to orthogonal directions in which nearby points are compressed. When the evolution is deterministic the geometry can become stretched to the point that the attractor occupies a measure 0 subset of Euclidean space. The probability distributions supported on this space are called singular, and are not in \mathfrak{F} , but if a little bit of randomness is added to the system then an invariant probability measure on the attractor might be in \mathfrak{F} , and have a local geometry characterized by stretching and compression in transverse directions.

The p -sphere, max-norm, and hyper-rectangular volume elements used by most k nn methods are not suitable for capturing this type of local geometry. These volume elements can be described as highly geometrically regular. This regularity serves a purpose in that it minimizes the amount of data needed to define the volume elements, and therefore allows the volume elements as local as possible. A drawback in data-driven applications where sample size is fixed and often limited is that the local volume elements might not be descriptive of the geometry of the underlying probability measures, resulting in bias in the estimators. A simple example of this problem is shown in Figure 3.2a, in which X and Y are normally distributed with standard deviation 1 and correlation $1 - \alpha$. By direct computation, the true mutual information increases asymptotically as $\log(\alpha)$, but for each k the KSG estimator

applied to the raw data diverges quickly as α decreases. Figure 3.2b illustrates the cause of the problem, which may be due to local volume elements not being descriptive of the geometry of the underlying measure. Improving on that issue is the major stepping off point of this section. In particular, the KSG local volume elements mostly resemble the green square (a max-norm sphere), whose volume greatly overestimates the volume spanned by the data points it contains.

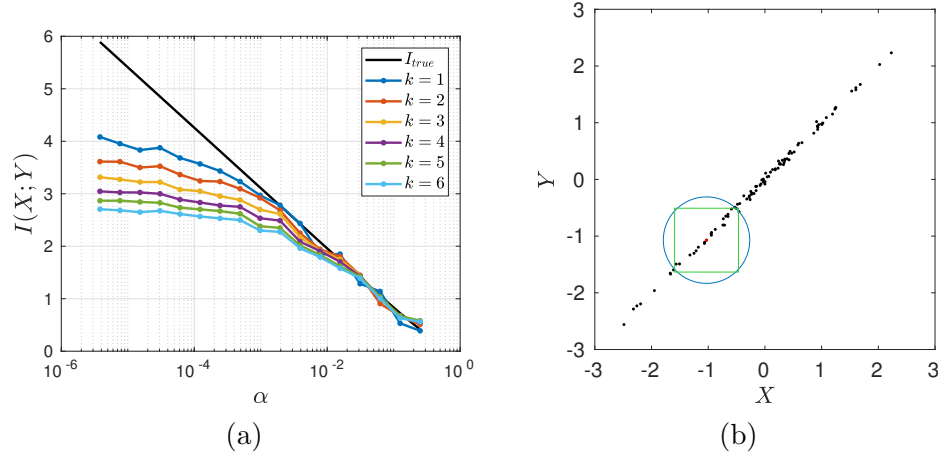


Figure 3.2: (a) KSG estimates of mutual information using max-norm spheres for two 1d normally distributed variables with standard deviation 1 and correlation $1 - \alpha$. For each $\alpha \in \{2^{-j} : j = 2, \dots, 18\}$ a sample of size $N = 100$ is drawn and the mutual information estimated by KSG with $k = 1, \dots, 6$. The true mutual information, $I_{true} = I(X; Y)$ is plotted in black. (b) A sample of size 100 when $\alpha = .001$. A randomly chosen sample point is highlighted in red. A sphere in the maximum norm is plotted in green and a sphere in the Euclidean norm is plotted in blue. The radius of each sphere is equal to the distance to the 20th closest neighbor in the respective norm.

This section introduces a new class of k nn estimators, the g-knn estimators, which use more irregular local volume elements that are more descriptive of the underlying geometry at the smallest length scales represented in the data. The defining feature of g-knn methods is a trade-off between the irregularity of the object, which requires more local data to fit, and therefore less localization of the volume elements, and the improvement in the approximation of the local geometry of the underlying measure.

Because of this trade-off, the local volume element should be chosen to reflect the geometric properties expected in the desired application. Motivated by the study of dynamical systems, it is reasonable to model these properties on the local geometry of attractors, which is characterized by stretching and compression in orthogonal directions.

To test the idea behind the g-knn estimators, Sec. 3.3.2 develops a particular g-knn method that uses local volume elements to match the geometry of stretching and compression in transverse directions. Ellipsoids are a good option for capturing this geometric feature because they have a number of orthogonal axes with different lengths. They are also fairly regular geometric objects: the only parameters that require fitting are the center and one axis for each dimension. Such ellipsoids can be fit very efficiently using the singular value decomposition (svd) of a matrix formed from the local data [45].

The g-knn estimator is tested on four one-parameter families of joint random variables in which the parameter controls the stretching of the geometry of the underlying measure. The estimates are compared with the KSG estimator as the local geometry of the joint distribution becomes more stretched. Distributions can also appear to be more stretched locally if local neighborhoods of data increase in size, which occurs in knn methods when sample size is decreased. Therefore, the g-knn estimator and KSG are also compared numerically in examples using small sample size.

Unlike the Kozachenko-Leonenko and KSG estimators, the g-knn method developed here has not been corrected for asymptotic bias, so that it should be expected that KSG outperforms this particular g-knn method for large sample size. What is surprising is that the g-knn estimator developed in Sec. 3.3.2 outperforms KSG for small sample sizes and thinly supported distributions despite lacking KSG's bias cancellation scheme. Since KSG is considered to be state-of-the-art in the nonparametric

estimation of mutual information, the result should hold for other methods that do not account for local geometric effects.

There have been many attempts to resolve the bias of KSG. For instance, Zhu *et al.* [144] improved on the bias of KSG by expanding the error in the estimate of the expected amount of data that lies in a local volume element. Also, Wozniak and Kruszewski [140] improved KSG by modeling deviations from local uniformity using the distribution of local volumes as k is varied. These improvements do not directly address the limitations of spheres to describe interesting features of the local geometry.

The class of g-knn estimators can be thought of as generalizing the estimator of mutual information described by Gao, Steeg, and Galstyan (GSG) in 2015 [40], which uses a principle component analysis of the local data to fit a hyper-rectangle. The svd-based g-knn estimator defined in Sec. 3.3.2 improves on the GSG treatment of local data. These improvements are highlighted in Sec. 3.3.2.

3.3.2 Method

This section defines a g-knn estimator of entropy that, in turn, yields an estimate of mutual information when substituted in Eq. (3.42). The g-knn estimator is based on resubstitution (See Def. 3.2.1).

Let $X = (X_1, \dots, X_n)$ be a sample of an \mathbb{R}^d -valued variable.

Svd estimation of volume elements

The elliptical local volume elements are estimated by singular value decomposition (svd) of the local data. For any fixed $i \in \{1, \dots, n\}$, denote the k nearest neighbors of X_i in \mathbb{R}^d by the random variables $X_{\rho_j(i)}$, for $j = 1, \dots, k$, where $\rho_j(i)$ is a $\{1, \dots, k\}$ -valued random variable defined so that $X_{\rho_j(i)}(\omega)$ is the j th closest point to $X_i(\omega)$ in

the set $\{X_p(\omega) : p \neq i\}$. The set $\{X_{\rho_j(i)} : j \in 0, \dots, k\}$, where $X_{\rho_0(i)} = X_i$ will be called the k -neighborhood of sample points of X_i .

In order for the svd to indicate directions of maximal stretching it is first necessary to center the data. Let $Z = \frac{1}{k+1} \sum_{j=0}^k X_{\rho_j(i)}$ be a random variable describing the centroid of the k -neighborhood in \mathbb{R}^d , and define the centered variables by $Y_i^j = X_{\rho_j(i)} - Z$.

In order for the svd, a matrix decomposition, to operate on the centered variables, define Y to be a $M_{(k+1),d}$ -valued variable, where in general, $M_{p,q}$ denotes the $p \times q$ real matrices, defined by

$$Y_i = \begin{pmatrix} Y_i^0 \\ Y_i^1 \\ \vdots \\ Y_i^k \end{pmatrix}. \quad (3.49)$$

Since $Y_i(\omega) \in M_{(k+1),d}$, it has an svd of the form $Y_i = U_i \Sigma_i V_i^T$, where U_i is a $U(k+1)$ -valued variable and $U(p)$ denotes the unitary matrices of size p , Σ_i takes values in the set of $(k+1) \times d$ real matrices which are zero with the possible exception of the nonnegative diagonal components, and V_i is a $U(d)$ -valued variable. The columns of U and V are \mathbb{R}^{k+1} and \mathbb{R}^d -valued random variables called the left and right singular vectors, and since the left singular vectors do not play a role in this estimator, the word “right” will be omitted when referring to the right singular vectors. The singular vectors will be denoted $V_i^{(l)}$. The diagonal components of Σ_i are \mathbb{R} -valued random variables, which will be denoted S_i^1, \dots, S_i^d .

Since V is unitary, the singular vectors, V_i^l are of unit length and mutually orthogonal, meaning for any $\omega \in \Omega$, $V_i^{l_1}(\omega) \perp V_i^{l_2}(\omega)$. The first singular vector, V_i^1 points in the direction in which the data is stretched the most, and each subsequent

singular vector points in the direction which is mutually orthogonal to all previous singular vectors and that accounts for the most stretching.

The singular values are equal to the square root of the sum of squares of the lengths of the projections of the Y_i^j onto a singular vector.

The use of data centered variables is an important difference between the g-knn estimator described here and the GSG estimator, in which the data is centered to X_i (see Footnote 2 in Ref. [40]). Centering the data at X_i can bias the direction of the singular vectors away from the directions implied by the underlying geometry. In Fig. 3.3, for instance, the underlying distribution from which the data is sampled can be described as constant along lines parallel to the diagonal $y = x$ and a bell curve in the orthogonal direction, with a single ridge along the line $y = x$. If local data were centered at the red data point then all vectors would have positive inner product with the vector $(-1, 1)$, so that the first singular vector would be biased toward $(-1, 1)$. The center of the local data, on the other hand, is near the top of the ridge, so that the singular vectors of the centered data (in blue in the figure) estimate the directions along and transverse to this ridge.

Translation and scaling of volume elements

Since the values of V_i^l in \mathbb{R}^d are orthogonal, the vectors $S_i^l V_i^l$ can be thought of as the axes of an ellipsoid centered at the origin. The ellipsoid needs to be translated to the k -neighborhood and scaled to fit the data. There are many ways to perform this translation and rescaling, three of which are depicted in Fig. 3.3 for a particular $\omega \in \Omega$.

In Fig. 3.3, there are two ellipsoids centered at the centroid of the k -neighborhood. The larger ellipsoid is the smallest ellipsoid that contains or intersects all points in the k -neighborhood. The problem with this approach is that these ellipsoids might

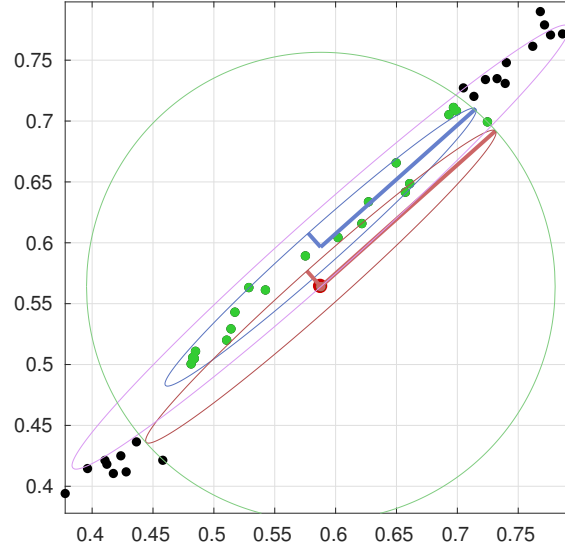


Figure 3.3: A data set, $X(\omega) = (x_1, \dots, x_n)$ sampled from a random variable with one sample point, x_i , highlighted in red at approximately $(0.59, 0.56)$, and its $k = 20$ nearest neighbors in Euclidean distance highlighted in green. The ellipsoid centered at x_i and drawn in red contains the volume used by the g-Knn estimator. The length of the major axes is determined by the largest projection of one of the k neighbors onto the major axes, enclosing $K_i(\omega) = 3$ points in its k -neighborhood, including itself. Two other ellipsoids are centered at the centroid of the $k + 1$ neighbors. The larger one (magenta) has radii large enough to enclose all $k + 1$ data points. The major axis of the smaller ellipsoid (blue) is determined by the largest projection of a data point onto the major axis. All three ellipsoids have the same ratio of lengths of axes, $S_1(\omega)/S_2(\omega)$, where the S_i are the singular values determined by the centered k -neighborhood.

contain data points which are not one of the k nearest neighbors of X_i , as is seen in Fig. 3.3. One solution to this problem is to include these data points in the calculation of the proportion of the data that lies inside the volume defined by the ellipsoid. We avoid this approach, however, both because it involves extra computational expense in finding these points, and because the new neighborhood obtained by adjoining these points is less localized.

An alternative approach is to decrease the size of the ellipsoid to exclude points not in the k -neighborhood, as depicted by the smaller ellipsoid centered at the centroid, Z . Such an ellipsoid could contain a proper subset of the k -neighborhood so that K_i , the \mathbb{Z}_+ -valued variable defined by the number of $X_j \neq X_i$ inside the ellipsoid, may be less than k . The problem with this approach, however, is that in higher dimensions, the ellipsoid may contain no data points, introducing a $\log(0)$ into Equation (3.34).

Instead of centering at the centroid, however, the ellipsoid could be centered at X_i . If the length of the major axis is taken to be the Euclidean distance to the furthest neighbor in the k -neighborhood, then the ellipsoid and its interior will only contain data points in the k -neighborhood because the distance from X_i to any point on the ellipsoid is less than or equal to the distance to the furthest of the k neighbors (the sphere in Fig. 3.3), which is less than or equal to the distance to any point not in the k -neighborhood. This neighborhood will contain at least one data point. An example of such an ellipsoid is shown in Fig. 3.3, which includes $K_i(\omega) = 3$ data points.

The GSG estimator [40] is centered at X_i , but the lengths of the sides of the hyper-rectangles seem to be determined by the largest projection of the local data onto the axes, which destroys the ratio of singular vectors that describes the local geometry. In addition, it is possible that the corners of the hyper-rectangles may circumscribe more than the k neighbors of X_i , even though a constant k neighbors are assumed in the estimate.

3.3.3 The g-knn estimates for $H(X)$ and $I(X; Y)$

In order to explicitly define the global estimate of entropy that results from this choice of center, define the \mathbb{R} -valued variable, $\epsilon_k(X_i)$, to be the Euclidean distance from X_i to the k th closest data point, $X_{\rho_k(i)}$. Define

$$R_i^l = \epsilon_k(X_i) \frac{S_i^l}{S_i^1} \quad (3.50)$$

to be the lengths of the axes of the ellipsoid centered at X_i , for $l = 1, \dots, d$. Note that $R_i^1 = \epsilon_k(X_i)$, and

$$\frac{R_i^{l_1}}{R_i^{l_2}} = \frac{S_i^{l_1}}{S_i^{l_2}}. \quad (3.51)$$

Letting v_d be the volume of a unit ball in \mathbb{R}^d , the volume of this ellipsoid can be determined from the formula

$$V_i = v_d \prod_{l=1}^d r_i^l \quad (3.52)$$

$$= v_d \epsilon_k(X_i)^d \frac{S_i^l}{S_i^1}. \quad (3.53)$$

Substitution into Equations (3.34) and (3.35) yields

$$\begin{aligned} \widehat{H}_{g-knn}(X) = & \log(n) + \log(v_d) - \frac{1}{n} \sum_{i=1}^n \log(K_i) + \frac{d}{n} \sum_{i=1}^N \log(\epsilon_k(X_i)) + \frac{1}{n} \sum_{i=1}^n \sum_{l=1}^d \log\left(\frac{S_i^l}{S_i^1}\right) \end{aligned} \quad (3.54)$$

The estimate for $I(X; Y)$ is then obtained using Eq. (3.42). The term $\frac{1}{N} \sum_{i=1}^N \sum_{l=1}^d \log\left(\frac{S_i^l}{S_i^1}\right)$ is small when the local geometry is relatively flat, but, as is demonstrated in Section 3.3.4, it can have a large impact on the estimate for more interesting local

geometries.

3.3.4 Examples

This section compares KSG estimates of mutual information on simulated examples with the estimates of the g-knn estimator defined in Sec 3.3.2. The examples are designed so that the local stretching of the distribution is controlled by a single scalar parameter, α . Plotting estimates against α suggests that the local stretching is a source of bias for KSG, but that the g-knn estimator is not greatly affected by the stretching.

The examples are divided into four one-parameter families of distributions in which the parameter α affects local geometry. Each family is defined by a model, consisting of the distributions of a set of variables and the equations that describe how these variables are combined to create X and Y . The objective is to estimate $I(X;Y)$ directly from a sample of size n without any knowledge of the form of the model.

The description of the families:

The first three families are designed to be simple enough that the mutual information can be computed analytically. Families 1 and 2 are 2d examples with 1d marginals built around the idea of sampling from a 1d manifold with noise in the transverse direction. The third family is a 4d joint distribution with 2d marginals. The fourth family is designed to be more typical of dynamical systems research and is defined by a pair of coupled Hénon maps with dynamically added noise. As opposed to the first three families, the fourth family is too complicated to find the true mutual information so the qualitative behaviors of the estimators are compared.

Family 1 : The model is

$$Y = X + \alpha V \quad (3.55)$$

$$X, V \text{ i.i.d. } \mathcal{U}(0, 1). \quad (3.56)$$

The result is a support that is a thin parallelogram around the diagonal $Y = X$. As $\alpha \rightarrow 0$ the distributions become more concentrated around the diagonal $Y = X$. The mutual information of X and Y is

$$I(X; Y) = -\log(\alpha) + \alpha - \log(2). \quad (3.57)$$

Family 2 : The second example is meant to capture the idea that noise usually has some kind of tail behavior. In this example V is a standard normal, so that the noise term, αV is normally distributed with standard deviation α . The model is

$$Y = X + \alpha V \quad (3.58)$$

$$X \sim \mathcal{U}(0, 1), \quad (3.59)$$

$$V \sim \mathcal{N}(0, 1), \quad (3.60)$$

where X and V are independent.

In this case the exact form of the mutual information is

$$I(X; Y) = -\log(\alpha) + \Phi\left(-\frac{y}{\alpha}\right) - \Phi\left(\frac{1-y}{\alpha}\right) - \frac{1}{2}\log(2\pi e), \quad (3.61)$$

where Φ is the cdf of the standard normal distribution.

Family 3 : In the third example the joint variable is distributed as

$$(X, Y) \sim \mathcal{N}(0, \Sigma) \quad (3.62)$$

$$\Sigma = \left[\begin{array}{cc|cc} 7 & -5 & -1 & -3 \\ -5 & 5 & -1 & 3 \\ \hline -1 & -1 & 3 & -1 \\ -3 & 3 & -1 & 2 + \alpha \end{array} \right]. \quad (3.63)$$

The first two coordinates of this variable belong to the variable X and the third and fourth to the variable Y . Thus, the upper left 2 by 2 block is the covariance matrix of X and the bottom 2 by 2 block is the covariance of Y . As long as $\alpha > 0$, Σ is positive definite but if $\alpha = 0$ then Σ is not of full rank, and the distribution $\mathcal{N}(0, \Sigma)$ is called degenerate, and is supported on a 3d hyperplane. When α is positive but small, the distribution can be considered to be concentrated near a 3d hyperplane. In this case the mutual information of X and Y is

$$I(X; Y) = -\frac{1}{2} \log \left(\frac{|\Sigma_X| |\Sigma_Y|}{|\Sigma|} \right) \quad (3.64)$$

where $|\cdot|$ is the determinant.

Family 4: The system is

$$X_{1,n+1} = a - X_{1,n}^2 + bX_{2,n} + \eta_1 \quad (3.65)$$

$$X_{2,n+1} = X_{1,n} + \eta_2 \quad (3.66)$$

$$Y_{1,n+1} = a - (cX_{1,n}Y_{1,n} + (1 - c)Y_{1,n}^2) + bY_{2,n} \quad (3.67)$$

$$Y_{2,n+1} = Y_{1,n} \quad (3.68)$$

$$\eta_1 \sim \mathcal{U}(-\alpha, \alpha) \quad (3.69)$$

$$\eta_2 \sim \mathcal{U}(-\alpha, \alpha), \quad (3.70)$$

where $a = 1.2$, $b = 0.3$, and the coupling coefficient is $c = 0.8$. When $\alpha = 0$ the system is the same coupled Hénon map described in a number of studies [69] except that a is reduced from the usual 1.4 to 1.2 so that noise can be added without causing trajectories to leave the basin of attraction. In these studies it is noted that a coupling coefficient of $c = 0.8$ results in identical synchronization so that in the limit of large n , $X_{1,n} = Y_{1,n}$ and $X_{2,n} = Y_{2,n}$, implying that the limit set is contained in a 2d manifold. Thus, in the long term, and as $\alpha \rightarrow 0$, samples from this stochastic process should lie near a 2d submanifold of \mathbb{R}^4 .

Families 1 through 3 are simple enough that, instead of using g-knn, one might be able to guess the algebraic form of the model and perform preprocessing to isolate noise and consequently remove much of the bias due to local geometry. For Families 1 and 2, for instance, if the algebraic form of the model $Y = X + \alpha V$ is known, one can express the mutual information as $I(X; Y) = H(Y) - H(\alpha V) + I(X; \alpha V)$, where the term $I(X; \alpha V)$ can be estimated by KSG after dividing by standard deviations (or, if αV is thought to be independent noise, then it would be assumed that $I(X; \alpha V) = 0$). This rearrangement of variables is in essence a global version of what the g-knn

estimator accomplishes locally using the svd.

A more complex example, which would likely resist efforts to preprocess the data to counteract the effects of local geometry, is provided by a 4d system consisting of coupled Hénon maps. In this system both X and Y are 2d and Y is coupled to X , but not vice-versa, so that X can be thought of as driving Y . The purely deterministic system approaches a measure 0 attractor so that $I(X; Y)$ would not be defined without the addition of noise, which is added to the X variable, and reaches the Y variable through the coupling.

It is important to note that the noise in Family 4 is introduced dynamically, and transformed by a nonlinear transformation on each time step. The samples lie near the Hénon attractor embedded in the 2d submanifold. If one thinks of the data as Hénon attractor plus noise, then the noise depends on X and Y in a nonlinear manner, and produces heterogeneous local geometries near each data point.

Numerical Tests

There are two ways in which the k -neighborhoods in these examples can get stretched. One way is that α gets small while n stays fixed. The other is that the local neighborhoods determined by the k nearest neighbors get larger. This occurs when α is small but fixed, and n is decreased, because the sample points become more spread out.

Choice of parameter k : Each of the g-knn estimates use a neighborhood size of $k = 20$. The value $k = 20$ was chosen because k should be small enough to be considered a local estimate, but large enough that the svd of the $(k + 1) \times d$ matrix of centered data should give good estimates of the directions and proportions of stretching. The value $k = 20$ is chosen because it seems to balance these criteria,

but no attempt has been made at optimizing k . Furthermore, in principle k should depend on the dimension of the joint space because the number of axes of an ellipsoid is equal to the dimension of the space. Therefore a better estimate might be obtained by using a larger k for Family 3 than the value of k used in Families 1 and 2.

For the KSG estimator, k is allowed to vary between 2 and 6. The value $k = 1$ is excluded because it is not used in practice due to its large variance. The most common choices of k are between 4 and 8 [41]. In each numerical example in this section, however, the KSG estimates become progressively worse as k increases, so that we omit the larger values of k in order to better present the best estimates of KSG.

Explanation of numerical results: Figures 3.4 and 3.5 show the results of KSG and g-knn estimates on samples of each of the four types of joint variables corresponding to the four families. In the figures on the left n is fixed at 10^4 and α varies. In the figures on the right n is allowed to vary but α is fixed at $1/100$. The top row of Figure 3.4 correspond to samples from Family 1, the middle row to Family 2, and the bottom row to Family 3. Figure 3.5 shows the results for Family 4. For each value of α or n , one sample of size n of the joint random variable is created and used by both estimators. In Fig. 3.4 the true value of the mutual information is plotted as a solid black line.

As thickness is varied: In each of Figs. 3.4a, 3.4c the true value of $I(X; Y)$ increases asymptotically as $\log(\alpha)$ when $\alpha \rightarrow 0$ and in Fig. 3.4e it increases as $\frac{1}{2} \log(\alpha)$. The KSG estimates do well for larger α but level off at a threshold that depends on the family. Although a bias is indicated by an average, since all of the KSG estimates beyond this threshold are less than the true value, it is safe to conclude that KSG becomes biased beyond this threshold. If the estimates were unbiased then by chance

around half of them would appear above the black line. The bias gets worse as α decreases, indicating that local geometry is a likely cause of the bias. The g-knn estimator, in contrast keeps increasing like $\log(\alpha)$ as α gets smaller, suggesting that the adaptations to traditional knn methods allow the g-knn method to adapt to the changing local geometry. It should be noted that the g-knn estimates in 3.4a also show signs of bias below the threshold where KSG becomes biased. An important difference, however, is that this bias does not seem to depend on α .

Figure 3.5a compares the estimates given by the g-knn estimator and the KSG estimator in Family 4 as α is decreased. The exact value of $I(X;Y)$ is unknown, but there are qualitative differences between the performance of the two estimators. Qualitatively, it is expected that as α decreases the true value of $I(X;Y)$ increases unboundedly. This behavior is captured by the g-knn estimator, which appears to increase asymptotically at about the rate of $1.95 \log(\alpha)$. The qualitative behavior is not captured by the KSG estimator, which appears asymptotically constant as $\alpha \rightarrow 0$.

In all four examples there is evidence that the variance may be relatively low for both the KSG and g-knn estimator. Although estimating the variance of the estimator would require many samples per value of α , the relative colinearity of the estimates for each estimator might be taken as evidence of low variance. Intuitively, a high variance would suggest that the estimates could go up or down a lot depending on the sample, and would therefore not lie in a straight line.

It is also interesting that the threshold beyond which KSG performs poorly occurs for larger values of α in Families 3 and 4 than in Families 1 and 2. The reason for this can be explained by the fact that Families 3 and 4 have 4d joint distributions. Colloquially speaking, there is a lot more space in higher dimensions, so the mismatch between the volume element and the local geometry can be much greater.

As sample size is varied: In each of Figs. 3.4b, 3.4d, 3.4f, and 3.5b, $\alpha = 10^2$ is fixed and n is allowed to vary. The true value is constant for all n since it is determined by the distribution, which depends on α alone. For large n , KSG outperforms g-knn since it is asymptotically unbiased [41]. In Fig. 3.4b, whereas KSG converges relatively rapidly, and is visibly indistinguishable from the true value by $n \approx 2 \cdot 10^5$, the g-knn estimator seems to have slightly negative asymptotic bias. In Fig. 3.4d the convergence of KSG seems even faster, and yet the g-knn estimator seems slightly negatively biased. In Fig. 3.5b the convergence of KSG is much slower (perhaps only linear for $n < 10^3$), which is likely due to the fact that the family has a 4d joint distribution. However, it does seem to have reached the true value by $n = 10^6$, whereas the g-knn estimates appear positively biased at $n = 10^6$ and may not have yet reached their asymptotic value. In Figure 3.5b none of the estimators seem to have leveled off by $n = 10^6$ ⁵.

The slow convergence of KSG in Families 3 and 4 suggests that the asymptotic convergence property may not be of much practical significance in application. Knowing that an estimator is asymptotically unbiased gives the hope that with a large data set the estimate should be near the true value. If $n = 10^3$ is a relatively large data set for a particular field of study, then Figs. 3.4f and 3.5b suggest that the KSG estimates could still be horribly biased. Furthermore, even if one could gather more data, they are faced with the computational time limits of KSG, which is at best $\mathcal{O}(n \log n)$. The examples suggest that increased dimension greatly slows down the convergence. In fact, the bias of KSG is $\mathcal{O}(n^{-1/d})$, where d is the joint dimension [41]. As these examples demonstrate, even for $d = 4$, and using simple examples that are

⁵The number $n = 10^6$ was chosen based on the computation time of KSG – higher n would have been unattainable in a reasonable amount of time. The g-knn computations are much faster than the KSG computations because the entropies are computed separately, and do not require a “range search” on each step.

defined by a simple set of equations, the estimator can be very biased. For more realistic and higher dimensional examples with as much local stretching and compression as the examples in Families 3 and 4, the KSG estimates would likely be extremely inaccurate.

The bias in KSG can be explained by a mismatch between the geometry of the local volume elements and the underlying measure. When n is small, the volume elements must be large in order to contain k elements. Such a volume element might not overlap much with the area of concentration of the underlying measure. In higher dimensions the volume element would have more volume which does not overlap the location where the measure is concentrated, and so there would be more bias, as indicated by Figs. 3.4f and 3.5b. As n gets larger, the sample points become denser, and in the limit of large n , perhaps even locally uniform. In this case, the KSG local volume elements are a good match for the local geometry and no bias results.

By using volume elements that match the local geometry, the g-knn estimator stays relatively constant under changes in n . For Families 1-3 the g-knn estimate stays near the true value as n decreases, and in Family 4 the g-knn estimates are stay relatively constant as compared to the KSG estimates. In Family 3 for low n , there are points both above and below the true value, giving the impression of an unbiased estimator. It is remarkable that in Families 1-3 the estimates stay relatively unbiased with as few as 32 sample points. It is possible that this is partly due to the linearity of the underlying measures (Families 1 and 2 lie close to a line and Family 3 lies close to a hyperplane). If the underlying measures were concentrated near a nonlinear manifold, then the local volume elements might not match the underlying measure as well. In particular, the major axis of the ellipsoid might behave like a tangent line segment to the underlying manifold, so that portions of the ellipsoid would not cover the manifold. This relates to the trade-off between having a less

local volume element (higher k) in exchange for greater description of the geometry of the underlying measure.

Although the bias remains relatively constant as n decreases, the variance seems to increase, as is indicated by the more jagged lines for the g-knn estimates for lower n in all four families. This might be expected because the ellipses are random ellipses, determined by the placement of the sample, and therefore, perturbations to the sample can alter the amount of overlap with the underlying probability measure. In most of these families, smaller sample size implies more elongated ellipses, so that a perturbation to the directions of the axes would have a larger effect. The effect could be much more complicated, however, and more research would need to be done. That said, in each of the first three families, the added variance of g-knn for small n seems relatively acceptable as compared to the large bias of KSG.

3.3.5 Discussion

A common strategy in nonparametric (e.g., knn) estimation of differential entropy and mutual information is to use local data to fit volume elements. The use of geometrically regular volume elements requires minimal local data, so that the volume elements remain as localized as possible. This section introduces the notion of a g-knn estimator, which uses slightly more data points to fit local volume elements in order to better model the local geometry of the underlying measure.

As an application, this section derives a g-knn estimator of mutual information, inspired by a consideration of the local geometry of dynamical systems attractors. A common feature of dissipative systems and systems with competing time scales is that their limit sets lie in a lower dimensional attractor or manifold. Locally the geometry is typically characterized by directions of maximal stretching and compression, which are described quantitatively by the Lyapunov spectrum. Ellipsoids are used for local

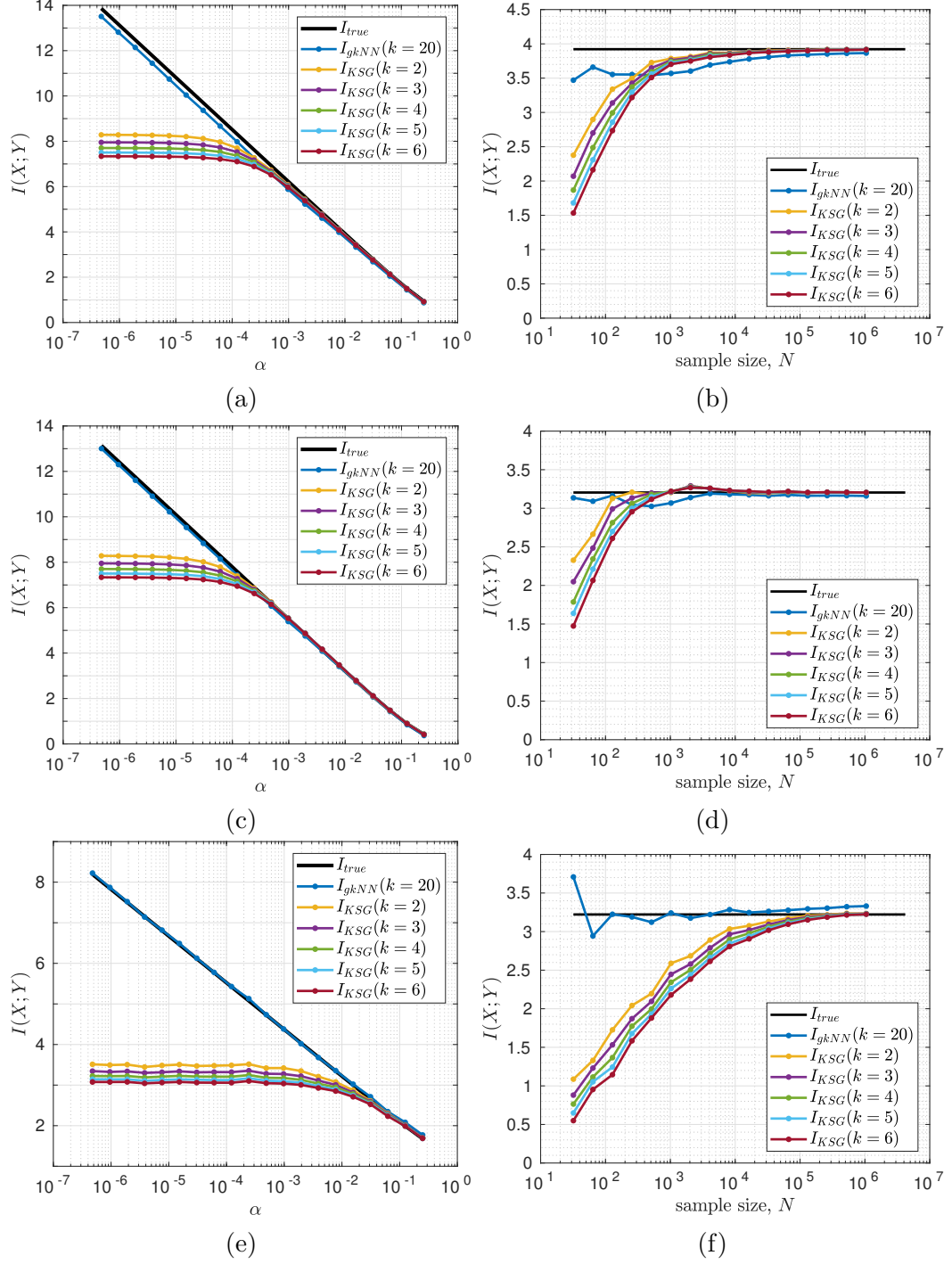


Figure 3.4: A comparison of the max-norm KSG estimator with the g-knn estimator on samples from the three families of variables. The top row of figures correspond to variables from family 1, the middle row to variables from family 2, and the bottom row to variables from family 3. In Figs. (a), (c), and (e), the sample size n is fixed at 10^4 and the thickness parameter α of each family is varied. On the right the thickness parameter of each family is fixed at $\alpha = .01$ and n is allowed to vary. For each value of α (left) or n (right) one sample of the joint random variable of size n is drawn and both the KSG and g-knn estimates are performed on the same sample.

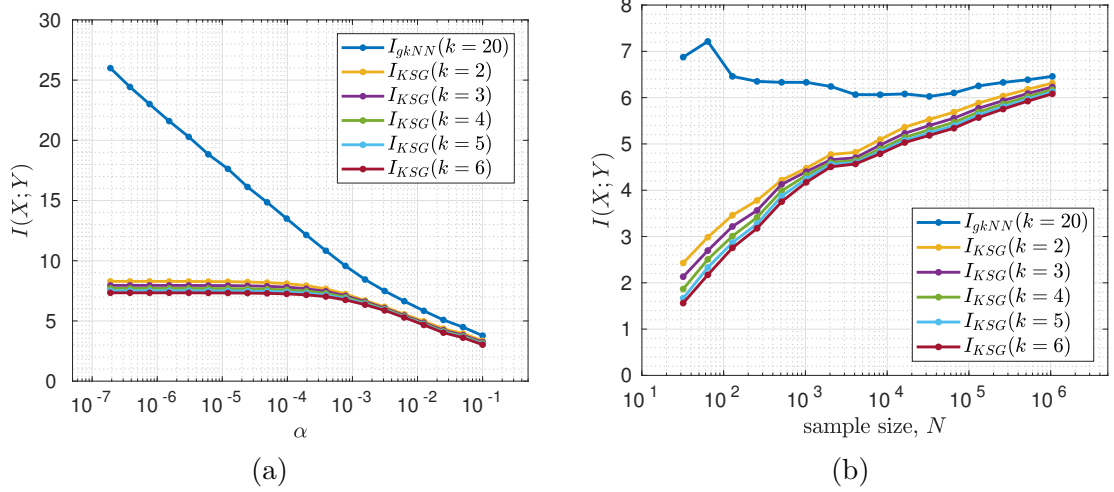


Figure 3.5: A comparison of the max-norm KSG estimator with the g-knn estimator for the stochastic coupled Hénon map described in Eqs.(3.65) through (3.70). In (a) the sample size is fixed at 10^4 while α is allowed to vary. In (b), α is fixed at $\alpha = .01$ and sample size varies.

volume elements because they capture the directions of stretching and compression without requiring large amounts of local data to fit.

It is worth noting that the ellipsoids are simply spheres in the Mahalanobis distance determined by the local data [77]. The metric that is used to define the spheres in the g-knn estimate, however, varies from neighborhood to neighborhood. This behavior is very different from many other knn estimators of pdfs where the spheres are determined by a global metric (typically defined by a p -norm). In this perspective, g-knn methods use data to learn both local metrics and volumes, and hence a local geometry, justifying the use of the name g-knn.

The numerical examples suggest that when it is not possible to preprocess the data to properly reflect the local geometry, the g-knn estimator outperforms KSG as the underlying measure becomes more thinly supported. The results suggest that the improvement to the bias might extend up to the boundary of \mathfrak{F} with the singular distributions. The g-knn estimator also outperforms KSG as sample size decreases,

a result that is particularly promising for applications in which the number of data points is limited. However, unlike the Kozachenko-Leonenko estimator of differential entropy and the KSG estimator of mutual information, the g-knn estimator as based on ellipses developed in this section is not asymptotically unbiased. In our future work we hope to eliminate this asymptotic bias in a manner analogous to Ref. [112] to gain greater accuracy for both low and high values of n .

There are also other descriptions of local geometry that suggest alternative g-knn methods. For instance, there are nonlinear partition algorithms such as OPTICS [4], which are based on data clustering. Also, entropy is closely related to recurrence, which is suggestive of alternative g-knn methods based on the detection of recurrence structures [46].

3.4 Nonparametric Hypothesis Testing

In addition to the estimation of CSE from data, a key step in the algorithmic inference (such as oCSE) of direct causal links from data is determining whether or not the estimated CSE value $\widehat{C_{X \rightarrow Y|Z}}$ should be regarded as being strictly positive.

This is an example of a one-sided hypothesis test. Hypothesis tests are much like estimators for estimands that can take only one of two values. Let g be an estimand and suppose that $\{A_0, A_1\}$ is a measurable partition of \mathbb{R} . Then g divides \mathfrak{F} into two sets of distributions,

$$H_0 = \{\nu \in \mathfrak{F} : g(\nu) \in A_0\}, \text{ and} \quad (3.71)$$

$$H_1 = \{\nu \in \mathfrak{F} : g(\nu) \in A_1\}. \quad (3.72)$$

Define the loss function to be the 0-1 loss function [111],

$$L(\nu, d) = \begin{cases} 0 & \mathcal{X}_{A_1}(g(\nu)) = \mathcal{X}_{A_1}(d) \\ 1 & \mathcal{X}_{A_1}(g(\nu)) \neq \mathcal{X}_{A_1}(d). \end{cases} \quad (3.73)$$

Then the risk of using T when the distribution is ν is

$$R(\nu, T) = \begin{cases} \mathbb{P}(T(X) \in A_1) & \nu \in H_0 \\ \mathbb{P}(T(X) \in A_0) & \nu \in H_1. \end{cases} \quad (3.74)$$

Even in simple cases there are no optimal estimators [111]. The most common approach, however, is to notice that the risk can be divided into two types of errors, type I, and type II corresponding to the first and second lines of Eq. (3.74), and to put priorities on avoiding each type of error. It is customary to place an upper limit on the probability of making a type I error, α , and then do whatever is possible to minimize type II errors.

In parametric situations one can often directly calculate the distribution of $g(\nu)$ under the assumption $\nu \in H_0$. The distribution of $g(\nu)$ can then be used to define a rejection region as a subset of \mathbb{R} that integrates to α under the null distribution. This sets up a test under which if $T(X)$ is in the rejection region then the hypothesis $\nu \in H_0$ is rejected.

The nonparametric case is more tricky because instead of knowing the distributions for each $\nu \in H_0$, one only has a single sample. To address this, we consider a shuffle test for the null hypothesis of $\widehat{C_{X \rightarrow Y|Z}} = 0$ [124] (also see Ref. [22, 131, 132]). Let

$$((X_1, Y_1, Z_1), (X_2, Y_2, Z_2), \dots, (X_n, Y_n, Z_n)) \quad (3.75)$$

be a sample (a $\Psi^{n(d_X+d_Y+d_Z)}$ -valued variable), where $1, \dots, n$ serve as discretized time indices. For any permutation, σ , of $\{1, \dots, n\}$, define $X_\sigma = (X_{\sigma(1)}, X_{\sigma(2)}, \dots, X_{\sigma(n)})$. For a typical σ there should be no pattern of dependence between $X_{\sigma(i)}$ and Y_{i+1} , so that

$$C_{X_\sigma \rightarrow Y|Z} = 0. \quad (3.76)$$

There are also a large number of permutations, so that a single $\omega \in \Omega$ can be used to infer a distribution for the estimates, $\widehat{C_{X_\sigma \rightarrow Y|Z}}$ by treating σ as a random variable sampled from a different probability space.

A reasonable test for $C_{X \rightarrow Y|Z} = 0$ is $\widehat{C_{X \rightarrow Y|Z}} \leq \widehat{C_{X_\sigma \rightarrow Y|Z}}$. A type I error would occur when $C_{X \rightarrow Y|Z} = 0$, but the test finds $C_{X \rightarrow Y|Z} > 0$. Thus, an estimator with a type I error of approximately α , is defined by rejecting H_0 if $\widehat{C_{X \rightarrow Y|Z}} > \widehat{C_{X_\sigma \rightarrow Y|Z}}$ more than $1 - \alpha$ times a set number of permutations that will be sampled.

Many hypothesis tests of the sort are executed during the oCSE algorithm. The inclusion/exclusion of a link typically requires multiple testings that are not necessarily independent, and as such, the α in a single significance test of CSE should not be interpreted as the significance level of the directed links inferred by the entire oCSE algorithm. The false positive rate for the entire algorithm is expected to be somewhat larger than α . However, numerical tests suggest that for a class of multivariate Gaussian distributions the false positive ratio is closely related to the α used in the individual tests [124]. Exact correspondence between α and the significance level of the inferred links in general remains an open challenge.

Chapter 4

Application to collective animal motion

Both to demonstrate the types of information that can be gleaned from the direct interaction networks inferred by the oCSE approach and to show that such networks are computable for real empirical data sets that contain noise and other non-idealities, we apply oCSE to empirical measurements of swarming insects. We have published the results in an article entitled “Inference of causal information flow in collective animal behavior” in *IEEE Transactions on Molecular, Biological and Multi-Scale Communications* [73]. Here, we briefly describe the experimental methodology, including the insect husbandry procedures and data acquisition system, and then show the results of the oCSE computation. These results enable us to compare and contrast spatially nearest neighbors with direct causal neighbors.

4.1 Background

Collectively interacting groups of social animals such as herds, schools, flocks, or crowds go by many names depending on the specific animal species. In all cases, they tend to display seemingly purposeful, coordinated group-level dynamics despite the apparent absence of leaders or directors. These coordinated group behaviors appear to emerge only from interactions between individuals, analogous to the manner in which macroscopic observables are determined by microscopic interactions in statistical physics. Thus, collective behavior has captivated a broad spectrum of researchers from many different disciplines [5, 6, 29, 43, 49, 54, 74, 80, 81, 86, 97, 101, 113, 115, 116, 120, 122, 142, 143].

Making the analogy with statistical physics more concrete, it is reasonable to suggest that a deep understanding of collective group motion may arise from three parallel pursuits. We can perform a macroscopic analysis, focusing on the observed group-level behavior such as the group morphology [91] or the material-like properties [85, 86, 125]; we can perform a microscopic analysis, determining the nature of the interactions between individuals [55, 61, 74, 97]; and we can study how the microscopic interactions scale up to give rise to the macroscopic properties [134].

The third of these goals—how the microscopic individual-to-individual interactions determine the macroscopic group behavior—has arguably received the most scientific attention to date, due to the availability of simple models of collective behavior that are easy to simulate on computers, such as the classic Reynolds [102], Vicsek [134], and Couzin [24] models. From these kinds of studies, a significant amount is known about the nature of the emergence of macroscopic patterns and ordering in active, collective systems [127]. But in arguing that such simple models accurately describe real animal behavior, one must implicitly make the assumption that the in-

teractions between individuals are correctly represented. Any model of interactions has two key and distinct components: a specification of the mathematical form of the interaction, and, more fundamentally, a choice as to *which* individuals interact. Given that it is difficult to extract the appropriate social interaction network from empirical measurements, models typically replace this hard-to-measure social network with the simple-to-define proximity network [89]. Thus, it is assumed that individuals interact only with other animals that are spatially nearby. No matter what species is involved, the answer to the question of whether interactions are generally limited to or dominated by spatial local neighbors has strong implications. Recently, for example, scientists studying networks have shown that introducing even a small number of long range interactions into a lattice can impart qualitative changes to the observed macroscopic behavior [66, 135]. Consequently, the question of whether flocks or swarms or herds also contain long range interactions between individuals may have important implications for the understanding of collective motion.

Efforts to move past the simple framework of assuming that the local spatial neighborhood of an individual dominates its behavior have been largely theoretical [15, 16], as it is challenging to extract the underlying interaction network from measured data. Empirical methods have often relied upon various types of correlational¹ (or other pairwise) time-series analysis [97], which by design only captures linear dependence and fail to detect the nonlinear relationships that are typical in real-world applications (See Chapter 2 for a more detailed analysis).

An alternative paradigm would be to use information theoretic methods that are capable of detecting nonlinear dependence. Chapter 2 contains a more detailed development of information theoretic methods. In this chapter we apply CSE and the

¹Unless otherwise noted, throughout the chapter we use “correlation” to mean the commonly adopted “Pearson linear correlation” in data analysis.

oCSE algorithm. In the application of inferring interactions among animals, oCSE requires knowledge only of the positions (or velocities or accelerations) of individuals in a group and is thus directly computable from empirical data. Because we define interactions via the information theoretic notion of the direct exchange of information as detected by uncertainty reduction, we need not make any assumptions about the spatial proximity of interacting individuals or the precise mathematical form of interaction. To demonstrate the unique utility of this oCSE network inference algorithm, we apply it to experimental measurements of the motion of individuals in mating swarms of the non-biting midge *Chironomus riparius*. In addition to showing the computability of the CSE in this data set, the oCSE approach clearly reveals that spatial proximity and interaction are not synonymous, suggesting that a deep understanding of collective behavior requires more subtle analysis of interactions than simple position-based proximity metrics.

4.2 Experimental Methods

Many different species of insects in the order Diptera exhibit swarming as a part of their mating ritual [33], and such swarms are a well studied, canonical example of collective behavior. Swarms are also an excellent model system for testing the oCSE algorithm: since swarms are internally disordered and show little overall pattern or correlation [84], it is difficult to tell by eye which individuals, if any, are interacting.

Here, we apply the oCSE algorithm to data collected from the observation of swarms of the non-biting midge *Chironomus riparius* under controlled laboratory conditions. The laboratory group of Nicholas Ouellette at Stanford University maintained the colonies of *Chironomus riparius* and performed all aspects of data collection. Details of the insect husbandry procedures and experimental protocols have

been reported in detail elsewhere [62, 96], so we describe them only briefly here. The breeding colony of midges is kept in a cubic enclosure measuring 91 cm on a side; temperature and humidity are controlled via laboratory climate-control systems. Midge larvae develop in 9 open tanks, each containing 7 L of oxygenated, dechlorinated water and a cellulose substrate into which the larvae can burrow. Adult midges live in the same enclosure, typically sitting on the floor or walls when they are not swarming. The entire enclosure is illuminated from above by a light source that provides 16 hours of light in each 24-hour period. When the light turns on and off, male midges spontaneously form swarms. Swarm nucleation is encouraged and the position of the swarm is controlled by the positioning of a “swarm marker” (here, a 32×32 cm piece of black cloth) placed on the floor of the enclosure. The number of midges participating in each swarming event is uncontrolled; swarms consisting of as few as one or two midges and as many as nearly 100 have been observed[98].

To quantify the kinematics of the midges’ flight patterns, the group reconstructs the time-resolved trajectory of each individual midge via automated optical particle tracking. The midge motion during swarming is recorded by three Point Grey Flea3 digital cameras, which capture 1 megapixel images at a rate of 100 frames per second (fast enough to resolve even the acceleration of the midges [62]). The three cameras are arranged in a horizontal plane outside the midge enclosure with angular separations of roughly 45° . Bright light can disrupt the natural swarming behavior of the midges; thus, the group illuminates them in the near infrared, which the midges cannot see but that the cameras can detect. In each 2D image on each camera, midge positions are determined by simple image segmentation followed by the computation of intensity-weighted centroids. These 2D positions were then combined together into 3D world coordinates via stereomatching, using a pinhole model for each camera and calibrating via Tsai’s method [128]. To match the individual 3D positions together

into trajectories, the group used a fully automated predictive particle-tracking method originally developed to study highly turbulent fluid flows [90]. Occasionally, tracks will be broken into partial segments, due to mistakes in stereoimaging or ambiguities in tracking; to join these segments together into long trajectories, the group used Xu’s method of re-tracking in a six-dimensional position-velocity space [141]. After tracks were constructed, accurate velocities and accelerations were computed by convolving the trajectories with a smoothing and differentiating kernel [98]. The final data set for each swarming event therefore consists of time series of the 3D position and its time derivatives for each midge.

4.3 Inferring Insect Interactions using oCSE: Choice of variables, parameters, and conditioning

After acquiring the experimental data, several decisions need to be made with respect to the choice of variables, parameters, and conditioning in the oCSE algorithm, in order to produce meaningful results that are interpretable.

The data sets used here contained empirical measurements from 126 distinct swarming events with varying numbers of participating individuals. For each swarm, time series of position, velocity, and acceleration were collected for each individual insect. We narrowed down the data by considering the collection of long (> 1 second) disjoint time intervals in which the corresponding data contains exactly the same number (≥ 5) of insects in each such interval. In other words we restricted our studies to data sets where the same “actors” were at play throughout the window of study.

Since the datasets include the spatial trajectories, the available variables include position, velocity, acceleration (all 3D), or any reasonable function of one or more of these, for example functions that project onto individual coordinate axes. We used

the 3D acceleration data as opposed to position or velocity, for two reasons. One is that accelerations are often interpreted as “social forces” in the animal motion literature [96], and so seemed the most suited form of data for investigating interactions. The other reason is that the acceleration showed less autocorrelation than either the position or velocity data.

Next, given that the motion of the insects is not stationary, we expect the causal influences to vary over time even within each experiment. For this reason, we apply oCSE to infer causal networks from data defined in relatively small time windows instead of the entire time span of each experiment. We choose the time window size to be 1 second, corresponding to 100 data samples (at a sampling frequency of 100Hz), which seems to be the minimal window size that produce relatively continuous-in-time causal networks.

In addition, we seek to infer causal influences from other midges *beyond* the influence of a midge to itself. To ensure that the self-influence is properly accounted for, we modified the starting point of the oCSE algorithm as described in Sec. 1.7 to be $\mathcal{Z} = \{Y\}$ (for a given midge Y), and always keep $Y \in \mathcal{Z}$ in both the discovery and removal phases.

Furthermore, the time-lag τ is chosen to be $\tau = 0.05$ seconds (5 time steps) based on biological considerations. In particular, we observed that [96] the midges tend to travel in a straight line for (usually less or equal to) 0.1 seconds before making a sudden acceleration over the next few frames, as if there were gathering and processing information during their straight flight before reacting. Fourier analysis reveals that the most important frequency for this acceleration is 1 acceleration per 0.1 seconds. Thus, the time lag should be smaller than 0.1 seconds in order to capture these accelerations. However, making τ too small reduces the amount of useful information comparing against noise. The choice of $\tau = 0.05$ seconds achieves

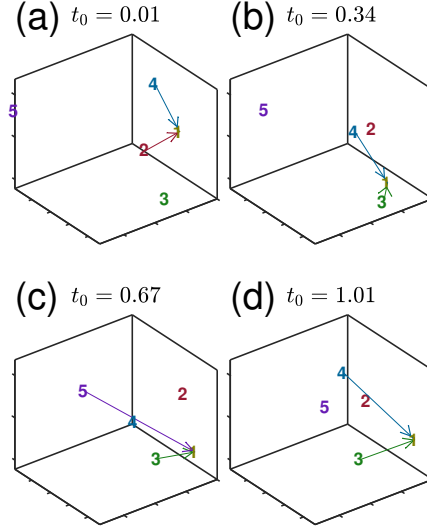


Figure 4.1: Information flow from the perspective of midge 1. Only information flows into midge 1 (i.e., direct causal links to 1) are depicted, and are represented by the solid lines. Each panel corresponds to an oCSE computation using 1 second (100 frames) of data. The positions of the midges are given by their initial positions, t_0 during the interval. The time lag—see the definition of CSE in Eq. (1.7.1)—is $\tau = 0.05$ seconds. The initial time $t_0 = 0.01$ of panel (a) is chosen to be the point when insect 5 becomes observable.

a reasonable compromise.

Finally, the estimator of CSE (see Appendix A) requires a choice of the parameter k . In this study we fixed $k = 4$ for all computations (too small of a k gives estimates with high variance), noting that a number of papers offer heuristics for choosing k [75, 117] as a function of sample size. We chose hypothesis tests with significance level $\alpha = 0.01$ in both the forward (discovery) and backward (removal) phases of the oCSE algorithm. In theory, α should control the sparsity of the desired graph but we were unable to confirm this numerically. We typically ran 1000 trials per hypothesis test.

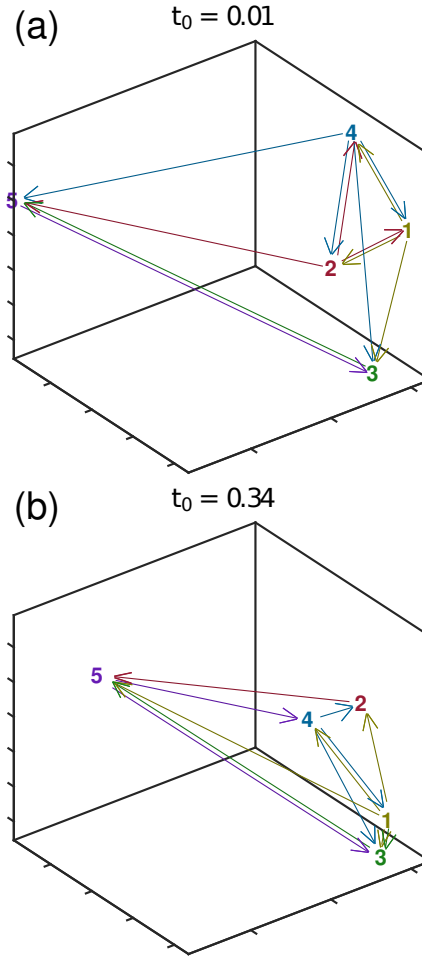


Figure 4.2: Directed graphs of all inferred information flows corresponding to the first two panels of Fig. 4.1. Each edge represents a flow of information and takes on the color of the information source. The parameters used in the oCSE algorithm are the same as in Fig. 4.1: $\tau = 0.05$, each computation uses 100 frames of data, and the positions are determined by the first frame.

4.4 Results

The most basic result of applying the oCSE algorithm is the determination of the *direct* causal links between individuals in the swarm. Although the data contains many suitable time series describing the motion of more than 30 swarming midges, Figs. 4.1 and 4.2 describe a small swarm of 5 midges for the purpose of illustrating the application of oCSE to finding direct causal links. In Fig. 4.1, we show four consecutive snapshots of these links from the perspective of a single insect (labelled as “1”) over a period of 2 seconds. In the first snapshot, panel (a), midge 1 is identified as being influenced by midges 2 and 4; that is, it is receiving information from them. Notice that in the second snapshot, (b), the link from 2 to 1 has been lost, but 1 is still receiving information from 4. Perhaps because it moved closer, 1 is also receiving information from 3 in the second snapshot. By panel (c), 1 seems to have noticed 5, but by the final snapshot this link has been lost.

Such transient interactions are reminiscent of those we described earlier using a different (time-series-based) measure [97]. In that case, we had hypothesized that the primary purpose of such interactions was for the registration of the gender of other midges in the swarm, since the biological purpose of swarming in this species is mating. A similar process may be at work here, and midge 1 may have, for example, successfully identified midge 2 after the first snapshot so that further information transfer was unnecessary.

In addition to studying only the information flows into a single insect, we can look at the entire directed graph returned by the oCSE algorithm. In Fig. 4.2, we show this full directed graph for the two initial conditions corresponding to frames (a) and (b) of Fig. 4.1. The direction of an edge is given by its color, where the source of the information flow determines the color.

One easy set of statistics to read off of these graphs are the in and out degrees. The in-degree of a node is the number of edges pointing to that node and the out-degree is the number of edges with one end at the node but pointing to a different node. In these plots, the out-degree of a node is the number of edges that are plotted in the same color as the node and the in-degree is the number of edges with an end at the node but which are plotted in a different color. So, for instance, in Fig. (4.2a), the in-degree of midge 1 is 2 (verifying the computation used to create Fig. (4.1a)), and the out-degree is 3.

The average in-degree (which is always equal to the average out-degree) is $12/5 = 2.4$ in both (a) and (b) of Fig. 4.2. In general, we found that the average number of causal neighbors per midge to be in between 2 and 3, and such number does not seem to change as a function of the swarm/network size in our experiments. In fact, for more than 95% of the cases over all analyzed swarming data (many of which contain > 5 midges), the maximum number of causal neighbors is always less or equal to 5. These findings suggest that a typical midge pays attention to a relatively constant number of other midges at any particular time. This hypothesis is lent more credence by noting that the in-degree of every individual midge is the same in (a) and (b).

The out-degrees are much more variable, however. Biologically speaking, out-degrees may give information on which midges are the most important, in the sense that if a midge has a high out-degree then others seem to be reacting to the motions of this midge. In Fig. (4.2a), although the most spatially central node, midge 2, has an out-degree of 3, so does midge 1, which is not as spatially central. Furthermore, midge 4 has the largest out-degree with every other midge paying attention to 4. So, although midge 2 is the most spatially central node, we say that midge 4 has the highest “degree centrality”. A similar analysis can be carried out on panel (b) showing that at $t_0 = 0.34$, midge 4 is now the most spatially central, but node 1

has the highest degree centrality. Some statistics that give more detailed information about centrality are eigenvector centrality and betweenness centrality.

Rather than attempting to comprehensively apply all available graph analysis methods, we give two simple observations and refer the reader to [83] for more on the analysis of graphs. In Fig. (4.2a) the subset of midges $\{3, 5\}$ forms a “sink” for information. Although 3 and 5 are gathering information from many midges, they are apparently unable to send that information to any other midges than 3 and 5. If one were to code edges in an “adjacency matrix” of 1’s representing edges and 0’s represent the lack of an edge, this feature corresponds to the adjacency matrix being reducible. The set $\{3, 5\}$ generates the only non-trivial subgraph closed under inclusion of all out-going edges. A less restrictive analysis that is similar in flavor is community detection [83], but this type of analysis is usually reserved for larger graphs.

Again in Fig. (4.2a), the edges linking $\{1, 2, 4\}$ generate a triangle in which information can flow in both directions around the triangle. This is a special relationship between three nodes called a 3-clique. It should be compared with $\{2, 4, 5\}$ in Fig. (4.2b) in which information flows only in one direction. In swarms with many other midges it is conceivable that most randomly picked triplets $\{a, b, c\}$ would have no triangle between them. The density of different types of triangles in a set is quantified by the clustering coefficient. Social networks tend to have much higher clustering as measured by clustering coefficients than technological networks and networks whose edges are determined randomly [82].

Because both Figs. 4.1 and 4.2 show that the configuration of causal links can and does change in time, it is reasonable to ask about the temporal variability and stability of the oCSE results: if the links switch seemingly at random from time step to time step, then the results would be unintelligibly unreliable. To check the

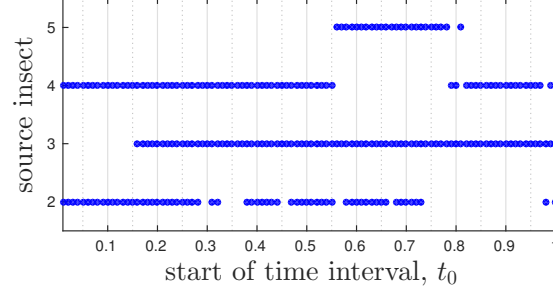


Figure 4.3: Evolution of a causal neighborhood of a single midge over one second following the trajectories of 5 midges. This figure gives a more (temporally) detailed look at the information flows depicted in Fig. 4.1 at the expense of spatial information. Again, midge 1 is the target individual; that is, we are inferring causal links toward midge 1. Edges from j to 1 are replaced by a point at (t_0, j) for the appropriate t_0 . The inputs and parameters are the same as those used to create Fig. 4.1. Long stretches of symbols or blank space demonstrate that the oCSE algorithm is robust to changes in the spatial configuration and kinematics of the swarm.

stability and reliability of the oCSE results, we computed the causal links for sets of overlapping time intervals, as shown in Fig. 4.3 for a particular example. Although there is occasionally some drop-out of links from one instant to the next, the overall results of the algorithm are clearly stable and more-or-less continuous in time; and we conjecture that the discontinuously dropped causal links could in fact be restored, if necessary, by improving the tracking accuracy or by some post-processing step or some combination of both.

Finally, as noted above, it is intriguing to note that midges connected via causal links are not always the spatially closest to each other. Although a full characterization and complete understanding of the distinction between spatial proximity and causal information flow is beyond the scope of the present paper, we can at least describe at a statistical and macroscopic level the difference between those by measuring the probability density functions (pdfs) of the distance between nearest spatial neighbors and nearest causal neighbors. These pdfs are plotted in Fig. 4.4 and show that as compared to simply calculating minimum distances for given time slices from the raw

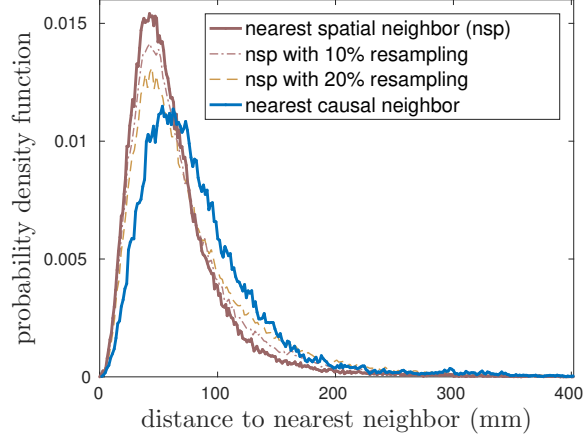


Figure 4.4: Probability density functions for distance from a midge to its: nearest spatial neighbor versus nearest causal neighbor. The distances used for constructing these pdfs are obtained by selecting a single midge randomly at each time slice and computing the minimal spatial distance to: other midges (spatial nearest neighbor distance), or causal neighbors. Furthermore, to account for the potential errors from statistical estimation and testing in causal inference, we also construct and plot the “resampled” pdfs of spatial nearest distances, by replacing a fixed fraction (10% or 20%) of these distances with distances to randomly chosen other midges.

data, the restriction to causal neighbors by the oCSE algorithm significantly shifts the typical distance between neighbors to larger values. Since choosing any neighbor other than the spatial nearest neighbor will always shift the distribution to the right, the nearest causal neighbor distribution should be compared to the distributions for nearest spatial neighbor in which 10% or 20% of the neighbors are chosen randomly. These distributions can be interpreted as the distribution of distances to neighbors under weakened nearest neighbor assumptions. So for instance, the distribution for distance to nearest spatial neighbor with 20% resampling is the distribution that would occur if each insect followed a rule of 4 out of 5 times following their closest neighbor and the remainder of times randomly following a random neighbor. The distribution of distances to nearest causal neighbors seems unlikely even under this loose interpretation of the nearest neighbor rule.

Chapter 5

Future directions

Understanding the behavior of complex systems, such as groups of animals and the human brain, is a fundamental challenge for the 21st century. Of particular interest is the notion of emergent behaviors. This thesis identifies interactions between components as fundamentally important for the understanding of emergent phenomena. The thesis focuses on the definition and inference of interactions in a complex system environment.

One of the most important contributions of the thesis is the demonstration of how CSE and oCSE expand the horizon of the types of scientific questions that are answerable using state-of-the-art mathematics. As an example, Chapter 4 applied oCSE to the question of whether the interactions between swarming insects of the species *Chironomus riparius* were limited to spatial neighbors. Only a few years ago, scientists might have called such a question mere “philosophical speculation,” because there would presumably never be a way to provide an answer using the tools of science. Section 5.1 discusses ongoing work that uses CSE and oCSE to further push the boundaries of what kinds of questions science can answer in an application of great scientific and societal importance – the human brain.

The scientific progress was made possible by the underlying mathematical ideas, beginning with the foundations of CSE, but leading to additional contributions to the math world. The primary examples are the exposition of the generalization of Shannon and differential entropy in Ch. 2, the presentation of the theory of estimation in Ch. 3, and the presentation of the g-knn class of estimators in Ch. 3. Section 5.2 suggests extensions of these ideas.

5.1 Application to brain function

Human cognition and brain function arises out of the coordination of many smaller scale brain parts, without any top-down coordination. There are many questions about this emergent behavior, such as what makes certain brains good at performing math or verbal tasks, or more prone to addiction, what causes psychiatric disorders, and how do brains regain functions after traumatic injuries, even when the injury has destroyed a part of the brain thought responsible for the given function? In order to answer these questions it is important to know which regions of the brain directly interact or share information during particular tasks.

The advent of fMRI and EEG technologies has made it possible to look inside an active living brain and observe areas of activity. A common use of these technologies is to observing which areas of the brain are active during specific cognitive tasks. The conclusions that can be drawn from this type of study are limited to observations that are somewhat static, such as “region A is used during cognitive task B.” Focusing on interactions introduces a new dimension to the type of question that can be asked. The new questions relate to the brain as a dynamic object as opposed to static. Instead of simply asking “what” questions, the focus on interactions allows the researcher to ask “how” questions.

5.1.1 Background

There are a few possible meanings regarding connectivity between two or more brain regions. Two regions might be considered connected if there are bundles of axons linking the regions. This is a type of physical connectivity that can be determined via dissection and staining of cells. This type of connectivity does not depend on the task that the brain is performing at the macroscale and therefore its structure alone cannot account for the ability to perform different tasks. Instead, neuroscientists are interested in brain functional connectivity (BFC), which is very loosely defined as a statistical relationship over time between activities at different sites in the brain [1, 2, 34, 78, 95, 103, 105].

The sites at which the activity is measured, which are the nodes in the BFC network, are generally taken to be specific physical areas in the brain, which are determined by the method used for collecting the data, but possibly aggregated to higher scales depending on the method being used. The most widely used method for collecting the data is functional Magnetic Resonance Imaging (fMRI), which measures a proxy for neuronal activity known as the blood-oxygen-dependent signal (BOLD) [39]. The oxygen in the blood is a fuel for the neurons, so BOLD measures the rate at which that fuel is being used. The fMRI divides the brain into about 20,000 spatial units called voxels. The number of voxels varies between subjects. The time lag between measurements is around two seconds. A less popular tool for discovering brain functional networks is electroencephalography (EEG), which makes more measurements in time but at the cost of spatial accuracy.

The nodes often consist of aggregations of voxels, often into anatomical regions. The regions of the brain are determined by physiological structure. The number of anatomical regions used in BFC studies has gradually increased over the years from 90 in the mid 2000's [2, 34, 105] to as many as 268 in 2016 [103].

The edges are typically defined in terms of correlations between activities at different sites. Ref. [114] reviews the definitions of edges that are commonly used in BFC. Some methods are as simple as a comparison of the correlation of activity at two different sites to a threshold value [34, 103]. This approach produces an undirected network, would only be able to detect linear relationships, and would not distinguish between directed and undirected relationships. A related method that is used for BFC is partial correlation [36, 105], which is similar to correlation with a threshold except that it conditions on other nodes, so that the edges can be interpreted as direct relationships (although still undirected).

Most of the results of these papers have focused on the large scale network structure. Some BFC networks have been found to be small world networks [1, 2], which are networks in which distances between the furthest nodes (as number of edges that need to be traversed to get from one node to the other) are low compared to the number of edges in the network. Other BFC networks have been found to be scale-free [2, 34], meaning that the degrees of the nodes (the number of neighbors a node has) appears as if they were sampled from a distribution with heavy tails. Yet other networks have been found to be modular [36], meaning that there exists communities of nodes with high numbers of edges within the community relative to the number of edges leaving the community. Directed edges have been defined using Granger causality [10].

5.1.2 Causation Entropy and new data sets

Each of the definitions of edges that are in common use does not fully capture what seems to be meant by a functional relationship. Most are not directional and the Granger causality method is limited to discovering linear relationships. A better definition of an edge is given by Causation Entropy (CSE) and the inference of the

network by the optimal Causation Entropy algorithm (oCSE).

To apply CSE and oCSE to analyzing brain activity, we are currently working with Paul Laurienti, a neuroscientist at Wake Forest, to apply CSE to fMRI data collected from subjects in his lab. Participants in Dr. Laurienti's research have been divided into a number of conditions. Two conditions consist of elderly subjects. In one of the conditions each subject performs a set of fMRI before and after undergoing a fitness program, and the subjects in the other condition do not do the fitness training. In the remaining groups of subjects the participants are asked to perform a number of cognitive tasks for a few minutes inside the fMRI machine. The tasks include resting, visualizing certain geometric objects, and an internet gambling task. In addition, Dr. Laurienti has the ability to run new subjects under a wide variety of conditions.

The data sets consist of fMRI data (BOLD levels) for around $n = 120$ samples in time and around 20,000 voxels depending on the subject. Each voxel is labeled with an anatomical region. The anatomical atlas that we are currently using has 116 regions. To reduce the number of dimensions in order to be able to perform the oCSE algorithm we are treating the nodes as average values of the BOLD level across each of the voxels in a single anatomical region. For some subjects, due to motion inside the machine, one of the smaller regions might not be recorded, so that each subject might have either 115 or 116 nodes.

These data sets offer a unique opportunity not only to infer a network but also to validate the use of CSE to define such networks. In particular, it should be possible to determine which cognitive task is being performed merely by analyzing the CSE network generated from the data. This can be thought of as a type of mind reading task. Consider a subject who performs each of two cognitive tasks. If one could use the data generated by each run to learn the CSE network associated with each task, then on subsequent runs it might be possible to determine which cognitive task the

subject was performing by using their fMRI data to produce a CSE network.

The fact that there will be error in the estimates of the CSE network (See the loss function of an estimator in Ch. 3) introduces a new mathematical challenge. The networks must be matched without knowing the exact networks. For this task it might be possible to generate a number of networks under each condition and then, given a new network, decide which group of networks it is most “similar to.” This challenge can be thought of as a classification problem. One way to approach the problem is to find an appropriate metric on the space of directed graphs. It is also possible to think of the classification problem as a hypothesis test, since the networks that are inferred by CSE are samples from a network-valued random variable defined by the estimators. However one thinks of it, this classification problem might arise frequently in analyzing the types of complex networks likely to be generated by fMRI data from human brains.

5.2 Entropy and estimation

One of the goals of the chapter on estimation (Ch. 3) is to show that estimation is an important and interesting area in mathematics, and yet, that certain areas, such as nonparametric estimation remain relatively unexplored. This paucity of exploration occurs in part because statistical estimation is a relatively new focus of serious mathematical inquiry, but also because, as the exposition demonstrates, it is inherently very challenging. Chapter 3 provides a novel account of the theory of estimation, presenting the theory as an ill-posed optimization problem in broad enough terms to include nonparametric estimation, and focusing on three strategies for ameliorating the ill-posed problem.

Chapter 3 shows that the strategy of weakening the Risk minimization condition

from uniform minimization to other forms, such as minimax, or Bayes risk, has important implications for knn estimators of differential entropy and related quantities. The chapter demonstrates that knn estimators have infinite minimax or Bayes risk over the set of distributions with finite entropy.

The problem arises when one gets too “close” to the boundary with the singular distributions. In the numerical examples, a parameter α is used to vary that “closeness.” But in order to extend numerical results to analytical findings, it would be useful to make the notion of “closeness” precise by defining a topology on the space, \mathfrak{F} . There are many metrics used for research on spaces of probability distributions [99], and the goal would be to find the most suitable one for describing “closeness” as we mean it – most likely in terms of the geometric aspects of the underlying measures.

As a further structural description of \mathfrak{F} , a Bayes risk requires a measure on \mathfrak{F} . In Chapter 3 it is noted that the part of \mathfrak{F} near the boundary with the singular distributions should be weighted strongly, because there are many examples of complex systems that are believed to produce distributions near this boundary. It would be desirable to define suitable Bayes risk, show that the Bayes risk of KSG and other knn methods is infinite, and to try to determine whether the Bayes risk of a g-knn estimator is finite.

Another strategy involves restriction to estimators with good properties. It is noted in the chapter that the g-knn estimator based on ellipsoids is not necessarily asymptotically unbiased. One strategy for removing any asymptotic bias is to directly calculate the asymptotic bias, and subtract the bias from the estimator. This is exactly how the Kozachenko-Leonenko estimator of differential entropy was created [67, 112]. Following these methods in parallel might actually be relatively straightforward – the main difference seems to be that an integral over $r > 0$ would be replaced by an iterated integral over r_1, r_2, \dots, r_d .

Another asymptotic property that would be interesting to study is convergence rates. The KSG convergence rate is $\mathcal{O}(n^{-1/d})$ [41], which means that increased dimension can greatly slow the rate of convergence, requiring more data to get a good estimate. Working in parallel to the analytical results that have been found for spherical volume elements might be a good place to start, but the convergence rate might also be established by simulation.

Other properties that can be explored by running many simulations per value of n include the bias and variance for small n . One issue not addressed by the numerical tests in Chapter 3 is what would happen if the underlying distributions were close to a curved manifold. Then it seems reasonable that there would be some bias as n decreased. One possibility is that the ellipsoids' major axes would act like a tangent plane, so that the ellipsoid would circumscribe more space away from the manifold. Another possibility is that near a sharp curve in the manifold, the ellipsoid would start looking more like a sphere.

The g-knn method based on ellipsoids is defined with the assumption that the underlying measure is close to a manifold. It would be interesting to see what happens when that assumption is broken, and try to fix it. One possibility is that the underlying “manifold” intersects itself. Near points of intersection the ellipsoids would look more like spheres, and fail to capture the underlying geometry. It is possible that a g-knn estimator that uses diffusion maps [23] to define the local volume elements would better find the local geometry.

It is also unclear how well the g-knn method based on ellipsoids would model the local geometry if the probability distribution were supported near a strange attractor. Thinking about a Lorenz attractor, a typical cross-section is very reminiscent of a Cantor set. With enough data sampled from a Lorenz attractor, one might hope that ellipsoids would become more and more stretched out along the individual trajectories

that form the Cantor set. But it is also possible that the addition of sample points in the direction of stretching from a given point on the attractor is matched by the addition of sample points in the transverse direction along the attractor. It would be interesting to see how well the g-knn estimator based on ellipsoids would perform, and if the ellipsoids do not describe the underlying geometry, then to derive a g-knn estimator that describes fractal geometries. A related question is, to what extent does a fractal plus a small amount of noise retain its fractal geometry, as opposed to acting like a neighborhood of a manifold?

Chapter 2 worked out in detail a generalized form of differential entropy based on a definition of KL-divergence which is different from the traditional definition used in information theory, but grounded in notions of absolute continuity and the Radon-Nikodym theorem. An interesting observation is that when the underlying distribution is highly symmetric, as is the case with a Haar measure, or the counting or uniform measure on a discrete set, then it might be possible to interpret the generalized differential entropy as measuring the relative lack of symmetry of the probability measure. This interpretation of entropy as measuring asymmetry is actually very close to the interpretation as uncertainty, or disorder, and it would be desirable to be able to make a definite connection between the purely measure-theoretic concept of entropy and notions from abstract algebra. It would be worthwhile to work out some examples to try to quantitatively connect differential entropy with properties of the groups (or semigroups) of symmetries of the probability measure.

This thesis focused almost primarily on interactions between \mathbb{R}^d -valued variables, but, the generality of the differential entropy defined in Chapter 2 makes it possible to talk about interactions between other types of variables. For example, there are collections of S^1 -valued random variables (often called oscillators), but the underlying Haar measure on the circle is simply a restricted Lebesgue measure, so it is possible

that the generalized entropy would not be needed. As Chapter 2 described, it is possible to define differential entropies with respect to a uniform measure on a strange attractor. This approach might be used to study interactions between components of dynamical systems that lie on measure 0 attractors. It is also possible to study unitary matrix-valued ($U(d)$) random variables with respect to Haar measure. In quantum information theory, for instance, there is the notion of a random quantum gate [35], and it is possible that these gates interact to create an emergent behavior. Unitary-valued random variables also show up in Chapter 3 in the definition of the g-knn estimator. The sample is a random variable, which defines a set of unitary and diagonal matrices (the svd), which are therefore also random variables, and the g-knn estimate (which is also a random variable) can be expressed entirely in terms of these unitary and diagonal matrices. In order to better understand the properties of the estimator, it might be useful to take advantage of the unitary-valued random variables. Since the estimator estimates entropy, it seems reasonable that the generalized differential entropy of the unitary-valued random variables could have some import.

These ideas are just a few examples of ideas that have been generated from the study and application of CSE and oCSE. They are indicative of a larger theme of science and math pushing forward together, one foot followed by the other. Before mathematical advancements such as CSE, many questions about the interactions within a complex system were thought to be out of the realm of science. The mathematical notion of components interacting by sharing information, as embodied by CSE, makes these questions more accessible, and it can be applied to swarming insects, the human brain, and an unlimited number of other complex systems. The act of pursuing these scientific advancements has lead to new exciting mathematical questions that have inspired the generalized notions of communication and informa-

tion described in Chapter 2, and the g-knn family of estimators in Chapter 3. Every successful solution of a mathematical or scientific problem has the potential to lead to exciting new problems, and it is hoped that the successes described in this thesis are able to inspire others and to further the tradition of math and science advancing together, step by step.

Appendices

Appendix A

Probability primer

This appendix defines random variables and distributions, and lists some of the consequences of these definitions. The definitions and results can be found in Refs. [14, 30, 50, 109, 145].

A.1 Random variables, distributions, and expectation

The real world is full of complicated systems in which it is difficult to predict outcomes. Even a system as simple as two dice bouncing across a table can be very difficult to model. Probability shifts the focus to considering sets of observables, or measurements, that can be performed.

Definition A.1.1 (Probability space and events). A probability space is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$ in which $\mathbb{P}(\Omega) = 1$. The elements of \mathcal{F} are called events.

Unless otherwise noted, all measures in this appendix are assumed to be positive measures. The measurable space (Ω, \mathcal{F}) should be thought of as large, since it represents all of the possible variations of the complex system under study. Therefore, it

will be assumed that \mathcal{F} is large enough to accommodate any definitions or structures that are needed.

In the rest of this appendix it is assumed that there is an underlying probability space, which is $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition A.1.2 (Random variable). Let (Ψ, \mathcal{A}) be a measurable space. Any measurable function

$$X : \Omega \rightarrow \Psi \tag{A.1}$$

is a Ψ -valued random variable, sometimes called a (Ψ, \mathcal{A}) -valued random variable for clarity.

The space Ψ is sometimes called the state space

Since (Ω, \mathcal{F}) can be large and unwieldy, it is desirable to use X to shift most of the analysis into the state space.

Definition A.1.3 (Distribution). Given a (Ψ, \mathcal{A}) -valued random variable, its distribution, ν , is the pushforward measure of \mathbb{P} onto Ψ . In particular, for any $A \in \mathcal{A}$,

$$\nu(A) = \mathbb{P}(X^{-1}(A)). \tag{A.2}$$

Probabilities of events in Ω such as $\mathbb{P}(\{\omega \in \Omega : X(\omega) \in A\})$, which is sometimes written $\mathbb{P}(X \in A)$, is easily translated to its equivalent form $\nu(A)$.

Integrals permit the notion of averaging a random variable according to the weights assigned to sets in \mathcal{F} by \mathbb{P} . For any statements involving integrals we will assume that the random variables are real-valued functions. One possibility is that $\Psi = \mathbb{R}$, so that X is real-valued. A more general situation is that the integrand is a real valued measurable function of Ψ .

Definition A.1.4 (Expected value). The expected value of X , if it exists¹, is

$$\mathbf{E}[X] = \int_{\Omega} X(\omega) d\mathbb{P}(\omega). \quad (\text{A.3})$$

If X has distribution ν on (Ψ, \mathcal{A}) , then

$$\mathbf{E}[X] = \int_{\Psi} x d\nu(x). \quad (\text{A.4})$$

Let g be a measurable real-valued function $g : \Psi \rightarrow \mathbb{R}$. Note that $g(X)$ is then a real-valued random variable on Ω , and

$$\mathbf{E}[g(X)] = \int_{\Omega} g(X)(\omega) d\mathbb{P}(\omega) \quad (\text{A.5})$$

$$= \int_{\Psi} g(x) d\nu(x). \quad (\text{A.6})$$

Let $\mathcal{L}^0(\Omega, \mathcal{F}, \mathbb{P})$ be the space² of \mathbb{R} -valued random variables. It is an algebra by pointwise operations (for instance $(aX + bY)(\omega) = aX(\omega) + bY(\omega)$). Define $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}) \subset \mathcal{L}^0(\Omega, \mathcal{F}, \mathbb{P})$ to be the \mathbb{R} -valued random variables with finite expectation. Then $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ is a sub-vector space of $\mathcal{L}^0(\Omega, \mathcal{F}, \mathbb{P})$ and \mathbf{E} is a real valued linear functional on $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$.

Definition A.1.5 (Variance). The variance of a random variable, $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, with expected value μ_X is defined by

$$\text{Var}[X] = \mathbf{E}[(X - \mu_X)^2] \quad (\text{A.7})$$

when the integral is finite. Otherwise the variance is sometimes said to be infinite.

¹The expected value might not exist if the integral is infinite

²A set for now, but it can be given some topological structure

Note that finite variance implies finite expected value. Define $\mathfrak{H} = \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P}) \subset \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ to be the space of variables with finite variance. The space \mathfrak{H} has an inner product defined by

$$\langle X, Y \rangle = \mathbf{E}[XY] \quad (\text{A.8})$$

and with this inner product \mathfrak{H} is a Hilbert space. If $\mathbf{E}[XY] = 0$ then X and Y are orthogonal. The quantity $\mathbf{E}[XY]$ is well-defined on much of $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P}) \times \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, and can even be extended to all of $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ by use of extended real values. In this sense the inner product extends the notion of orthogonality on \mathfrak{H} to $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$.

Usually one is interested in the similarities between X and Y that is not due to a constant term, and so the variables are often “centered” before applying the inner product.

Definition A.1.6 (Covariance and correlation). Let $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, with expectations μ_X and μ_Y . Then their covariance is defined by

$$\text{Cov}[X, Y] = \mathbf{E}[(X - \mu_X)(Y - \mu_Y)], \quad (\text{A.9})$$

if the integral is finite. If the variables have finite variances then the Pearson correlation of X and Y is defined by

$$\text{Cor}[X, Y] = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}[X]\text{Var}[Y]}}. \quad (\text{A.10})$$

A.2 Sub- σ -algebras, and product measures

Definition A.2.1 (Vector-valued random variables). If X is an \mathbb{R}^d -valued random variable (in other words, $\Psi = \mathbb{R}^d$, where the σ -algebra of Ψ is a Borel σ algebra

with respect to the standard topology on \mathbb{R}^d) then it is called a vector-valued random variable.

The expectation can be extended to vector-valued random variables by applying the univariate expectation component-wise, so that \mathbf{E} is an \mathbb{R}^d -valued functional,

$$\mathbf{E}[X] = (\mathbf{E}[X^1], \dots, \mathbf{E}[X^d]). \quad (\text{A.11})$$

Definition A.2.2 (σ -algebra generated by a random variable). Let X be a (Ψ, \mathcal{A}) -valued random variable. Define the σ -algebra generated by X to be

$$\sigma(X) = \{X^{-1}(A) : A \in \mathcal{A}\}. \quad (\text{A.12})$$

The σ -algebra $\sigma(X)$ is the smallest sub- σ -algebra of \mathcal{F} that makes X measurable.

Proposition A.2.1 (Vector-valued random variables are vectors of \mathbb{R} -valued random variables). *The random variable X is an \mathbb{R}^d -valued random variable if and only if there are \mathbb{R} -valued random variables X^i , $i = 1, \dots, d$ such that*

$$X = (X^1, \dots, X^d). \quad (\text{A.13})$$

The key to the proof is showing that $\sigma(X)$ is the smallest σ -algebra that makes all of the variables X^i , $i = 1, \dots, d$ measurable.

If X and Y are \mathbb{R}^d -valued random variables, then the covariance can be defined by treating $X(\omega)$ and $Y(\omega)$ as column vectors and defining

$$\text{Cov}[X, Y] = \mathbf{E}[(X - \mu_X)(Y - \mu_Y)^T], \quad (\text{A.14})$$

where the expectation is computed component-wise on the $d \times d$ matrix as if it were

a vector in \mathbb{R}^{d^2} .

Independence of X and Y is related to the distribution of the vector-valued random variable (X, Y) . For this we must first introduce product measure. This treatment of product measure is taken from Halmos' textbook on measure theory [51].

Definition A.2.3 (Product σ -algebra). Let (Ψ_1, \mathcal{A}_1) and (Ψ_2, \mathcal{A}_2) be measurable spaces. The product σ -algebra, $\mathcal{A}_1 \times \mathcal{A}_2$ is the smallest σ algebra on $\Psi_1 \times \Psi_2$ containing all sets of the form $A_1 \times A_2$ for $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$. That is,

$$\mathcal{A}_1 \times \mathcal{A}_2 = \sigma(\{A_1 \times A_2 : A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2\}). \quad (\text{A.15})$$

Definition A.2.4 (Product measure). Let $(\Psi_1, \mathcal{A}_1, \nu_1)$ and $(\Psi_2, \mathcal{A}_2, \nu_2)$ be measure spaces. Any measure $\nu_1 \times \nu_2$ on $(\Psi_1 \times \Psi_2, \mathcal{A}_1 \times \mathcal{A}_2)$ such that

$$\nu_1 \times \nu_2(A_1 \times A_2) = \nu_1(A_1)\nu_2(A_2) \quad (\text{A.16})$$

for all $A_1 \in \mathcal{A}_1$ and $A_2 \in \mathcal{A}_2$ is called a product measure.

Existence and uniqueness of such a measure is due to Carathéodory's extension theorem. The uniqueness aspect requires that the measure space be σ -finite.

Definition A.2.5 (Finite and σ -finite). The measure space (Ψ, \mathcal{A}, ν) is called finite if $\nu(\Psi) < \infty$. It is called σ -finite if there exists a sequence $\{A_i\}_{i \in \mathbb{N}}$ of elements of \mathcal{A} such that

$$1. \quad \Psi = \bigcup_{i \in \mathbb{N}} A_i$$

$$2. \quad \nu(A_i) < \infty.$$

The spaces \mathbb{R} and \mathbb{R}^d are both σ -finite. A probability space is finite.

Theorem A.2.2 (Existence and uniqueness conditions for product measure). *If $(\Psi_1, \mathcal{A}_1, \nu_1)$ and $(\Psi_2, \mathcal{A}_2, \nu_2)$ are σ -finite measure spaces then there exists a unique product measure, $\nu_1 \times \nu_2$, on $\mathcal{A}_1 \times \mathcal{A}_2$.*

This product measure can be written as an integral.

Definition A.2.6 (Sections). Let $E \subset \Psi_1 \times \Psi_2$. Let $x \in \Psi_1$ and $y \in \Psi_2$. The section determined by x is the subset of Y defined by

$$E_x = \{y : (x, y) \in E\} \subset Y. \quad (\text{A.17})$$

Likewise, the section determined by y is the subset of X defined by

$$E^y = \{x : (x, y) \in E\} \subset X. \quad (\text{A.18})$$

Proposition A.2.3. *Let $(\Psi_1, \mathcal{A}_1, \nu_1)$ and $(\Psi_2, \mathcal{A}_2, \nu_2)$ be σ -finite measure spaces, and $\nu_1 \times \nu_2$ be the product measure on $\Psi_1 \times \Psi_2$. Then $E_x \subset \mathcal{A}_2$, $E^y \subset \mathcal{A}_1$, and the functions $f(x) = \nu_2(E_x)$ and $g(y) = \nu_1(E^y)$ are measurable. Furthermore, for any $E \in \mathcal{A}_1 \times \mathcal{A}_2$,*

$$\nu_1 \times \nu_2(E) = \int_{\Psi_1} \nu_2(E_x) d\nu_1 = \int_{\Psi_2} \nu_1(E^y) d\nu_2. \quad (\text{A.19})$$

Note that if $g : \Psi_1 \times \Psi_2 \rightarrow \mathbb{R}$ is a real-valued measurable function and X is a Ψ_1 -valued random variable and Y is a Ψ_2 -valued random variable then $g(X, Y)$ is an \mathbb{R} -valued random variable and

$$\mathbf{E}[g(X, Y)] = \int_{\Psi_1 \times \Psi_2} g(x, y) d(\nu_1 \times \nu_2)(x, y). \quad (\text{A.20})$$

Theorem A.2.4 (Fubini's theorem). *Let $(\Psi_1, \mathcal{A}_1, \nu_1)$ and $(\Psi_2, \mathcal{A}_2, \nu_2)$ be σ -finite measure spaces. If $\int_{\Psi_1 \times \Psi_2} |g(x, y)| d(\nu_1 \times \nu_2)(x, y) < \infty$ or if g is nonnegative and mea-*

surable, then the functions $g_1(x) = \int_{\Psi_2} g(x, y) d\nu_2(y)$ and $g_2(y) = \int_{\Psi_1} g(x, y) d\nu_1(x)$ are integrable and

$$\int_{\Psi_1 \times \Psi_2} g(x, y) d(\nu_1 \times \nu_2)(x, y) = \int g_1(x) d\nu_1(x) = \int g_2(y) d\nu_2(y). \quad (\text{A.21})$$

Independence relates to product measures.

Definition A.2.7 (Independence). Two sub- σ -algebras $\mathcal{G}, \mathcal{H} \subset \mathcal{F}$ are independent if

$$\mathbb{P}(G \cap H) = \mathbb{P}(G)\mathbb{P}(H) \quad \forall G \in \mathcal{G}, H \in \mathcal{H}. \quad (\text{A.22})$$

Two random variables, X and Y are independent if $\sigma(X)$ and $\sigma(Y)$ are independent. The independence of X and Y is denoted $X \perp\!\!\!\perp Y$.

Proposition A.2.5 (Independence and factorization of distribution). *Let X and Y be \mathbb{R}^d -valued random variables with distributions ν_X and ν_Y and joint distribution $\nu_{X,Y}$ (on \mathbb{R}^{2d}). Then*

$$X \perp\!\!\!\perp Y \iff \nu_{X,Y} = \nu_X \times \nu_Y. \quad (\text{A.23})$$

The proof uses the definition of the σ algebra for the joint distribution and the fact that $\nu_1 \times \nu_2(A_1 \times A_2) = \nu_1(A_1)\nu_2(A_2)$.

If $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$ then independence implies that the variables are uncorrelated: $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$. It implies further that for any real-valued measurable functions g_1 and g_2 , that $g_1(X) \perp\!\!\!\perp g_2(Y)$. Hence, if X and Y are centered variables (0 expectation) then independence can be thought of as a strong form of orthogonality, which perhaps justifies the use of the symbol $\perp\!\!\!\perp$.

A.3 Absolute continuity, Radon Nikodym, and conditioning

There are potentially many measures defined on the range space (Ψ, \mathcal{A}) . In fact every random variable pushes forward a measure from $(\Omega, \mathcal{F}, \mathbb{P})$. Also, if ξ is a measure in (Ψ, \mathcal{A}) , then every measurable function $g : \Psi \rightarrow [0, \infty]$ induces a measure on (Ψ, \mathcal{A}) via

$$\nu_g(A) = \int_A g d\xi. \quad (\text{A.24})$$

The notion of absolute continuity establishes some structure on the measures on Ψ and the Radon-Nikodym uses that structure to create a partial converse to Eq. (A.24).

Definition A.3.1 (Absolute continuity). Let (Ψ, \mathcal{A}) be a measurable space and let μ and ν be measures on (Ψ, \mathcal{A}) . Then μ is absolutely continuous with respect to ν , written

$$\mu \ll \nu, \quad (\text{A.25})$$

if for all $A \in \mathcal{A}$,

$$\nu(A) = 0 \implies \mu(A) = 0. \quad (\text{A.26})$$

Theorem A.3.1 (Radon-Nikodym theorem). *Let μ and ν be σ -finite measures on a measurable space (Ψ, \mathcal{A}) such that*

$$\mu \ll \nu. \quad (\text{A.27})$$

Then there is an \mathcal{A} -measurable function $g : \Psi \rightarrow [0, \infty]$, such that for all $A \in \mathcal{A}$,

$$\mu(A) = \int_A g \, d\nu. \quad (\text{A.28})$$

The function g will be called the Radon-Nikodym derivative of μ with respect to ν and is denoted $\frac{d\mu}{d\nu}$.

See Sec. A.6 for more discussion and a proof of this important theorem.

In Prop. A.2.5 it is shown that independence implies that distributions factor into product measures. This type of factorization extends to Radon-Nikodym derivatives.

Proposition A.3.2. *Let $(\Psi_1, \mathcal{A}_1, \xi_1)$ and $(\Psi_2, \mathcal{A}_2, \xi_2)$ be σ -finite measure spaces and suppose $\nu_1 \ll \xi_1$, and $\nu_2 \ll \xi_2$. Then*

$$\frac{d(\nu_1 \times \nu_2)}{d(\xi_1 \times \xi_2)}(x_1, x_2) = \frac{d\nu_1}{d\xi_1}(x_1) \frac{d\nu_2}{d\xi_2}(x_2). \quad (\text{A.29})$$

The proposition can be proved by considering the sets in $\mathcal{A}_1 \times \mathcal{A}_2$ for which

$$(\nu_1 \times \nu_2)(A) = \int_A \frac{d\nu_1}{d\xi_1} \frac{d\nu_2}{d\xi_2} \, d(\xi_1 \times \xi_2). \quad (\text{A.30})$$

Fubini's theorem shows that this set includes all sets of the form $A_1 \times A_2$. More advance methods not covered in this appendix can be used to extend this class of measurable sets to $\mathcal{A}_1 \times \mathcal{A}_2$.

An important consequence of the Radon-Nikodym theorem is the existence of conditional expectation. However, a slight generalization of the Radon-Nikodym theorem as stated in Theorem A.3.1 is required. Although all other measures in the appendix are assumed to be positive measures (functions from \mathcal{F} to $[0, \infty]$), Theorem A.3.1 also holds even if μ is a signed measure (takes values in $[-\infty, \infty]$) and g in Eq. A.28 is allowed to also take values in $[-\infty, \infty]$.

Definition A.3.2 (Conditional Expectation). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, and let \mathcal{G} be a sub- σ -algebra of \mathcal{F} . The variable Z is a conditional expectation of X with respect to \mathcal{G} , denoted $Z = \mathbf{E}[X|\mathcal{G}]$ if

1. $Z \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$.
2. Z is \mathcal{G} -measurable.
- 3.

$$\int_A Z \, d\mathbb{P} = \int_A X \, d\mathbb{P}, \quad \forall A \in \mathcal{G}. \quad (\text{A.31})$$

The terminology “a conditional expectation” is used because there could be many such Z . It is not too hard to show that if Z and Z' are both conditional expectations of X given \mathcal{G} then $Z = Z'$ almost surely (they differ at most on a set of measure 0).

The existence of a conditional expectation can be established by the Radon-Nikodym theorem because the *signed* measure on \mathcal{G} defined by

$$\nu(A) = \int_A X \, d\mathbb{P} \quad (\text{A.32})$$

is absolutely continuous with respect to \mathbb{P} , so that

$$\int_A \frac{d\nu}{d\mathbb{P}} \, d\mathbb{P} = \nu(A) \quad (\text{A.33})$$

$$= \int_A X \, d\mathbb{P}, \quad (\text{A.34})$$

and the other two conditions on $\frac{d\nu}{d\mathbb{P}}$ follow from the construction.

In the Hilbert space $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ it is not too hard to show that the operator $X \mapsto \mathbf{E}[X|\mathcal{G}]$ is a projection. The subspace of \mathcal{G} -measurable random variables is a

linear subspace, closed under L^2 -norm, and $\mathbf{E}[X|\mathcal{G}]$ is the projection of X onto this subspace. It can be shown that this notion of conditioning as expectation extends into the space $\mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, and that $\mathbf{E}[X|\mathcal{G}]$ is the closest \mathcal{G} -measurable random variable to X in an L^2 -norm sense.

Definition A.3.3 (Conditioning with respect to a variable). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $X, Y \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$. Then the conditional expectation of X given Y is defined by

$$\mathbf{E}[X|Y] = \mathbf{E}[X|\sigma(Y)]. \quad (\text{A.35})$$

Definition A.3.4 (Conditional probability distribution). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let X be a (Ψ, \mathcal{A}) -valued random variable, and let $\mathcal{G} \subset \mathcal{F}$ be a sub- σ -algebra. A regular conditional probability distribution of X given \mathcal{G} is a function $\mathbb{P}_{X|\mathcal{G}} : \mathcal{F} \times \Psi \rightarrow [0, 1]$ such that

1. For a fixed $A \in \mathcal{F}$, $\mathbb{P}_{X|\mathcal{G}}(A, \cdot) = \mathbf{E}[\mathcal{X}_A X|\mathcal{G}]$ almost surely.
2. For a fixed $x \in \Psi$, $\mathbb{P}_{X|\mathcal{G}}(\cdot, x)$ is a probability measure on (Ψ, \mathcal{A}) .

The regular conditional probability distribution of X given a variable Y is given by substitution $\sigma(Y)$ for \mathcal{G} . In this case the probability measure in condition 2. will sometimes be written $\mathbb{P}(Y|X = x)$.

Definition A.3.5 (Support of a measure). Let (Ψ, \mathcal{A}, ν) be a measure space such that Ψ has a topology τ and \mathcal{A} is a Borel measure, meaning $\tau \subset \mathcal{A}$, then the support of ν is the set of all points, x , in Ψ such that every open set containing x has positive measure:

$$\text{supp } \nu = \{x \in \Psi : x \in U \in \tau \implies \nu(U) > 0\}. \quad (\text{A.36})$$

Definition A.3.6 (Probability density function). Let $\Psi = \mathbb{R}^d$ with Lebesgue measure, λ . If X is a \mathbb{R}^d -valued random variable with probability distribution ν , and

$$\nu \ll \lambda, \quad (\text{A.37})$$

then f_X is the probability density function of X if

$$f_X = \frac{d\nu}{d\lambda}. \quad (\text{A.38})$$

Definition A.3.7 (Cumulative distribution function). Let $\Psi = \mathbb{R}^d$ with Lebesgue measure, λ . If $X = (X_1, \dots, X_d)$ is a \mathbb{R}^d -valued random variable with probability distribution ν then the cumulative density function of X is a function $F_X : \mathbb{R}^d \rightarrow [0, 1]$ such that

$$F_X(x_1, \dots, x_d) = \mathbb{P}(X_1 \leq x_1, \dots, X_d \leq x_d). \quad (\text{A.39})$$

Theorem A.3.3. *Let $\Psi = \mathbb{R}^d$ with Lebesgue measure, λ , and let X be an \mathbb{R}^d -valued random variable with probability distribution ν , and*

$$\nu \ll \lambda. \quad (\text{A.40})$$

Then the probability density function, f_X and the cumulative distribution function, F_X are related by

$$f_X(x) = \frac{\partial}{\partial x_1 \cdots \partial x_d} F(x). \quad (\text{A.41})$$

A.4 Convergence of random variables

All variables in this section are assumed to be real-valued, but the definitions could be extended to the case where the variables are Ψ -valued and Ψ is a metric space.

Definition A.4.1 (Convergence in Probability). A sequence of random variables $\{X_n\}_{n \in \mathbb{N}}$ converges in probability to a random variable X if for all $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \epsilon) = 0. \quad (\text{A.42})$$

Convergence in probability is denoted $X_n \xrightarrow{p} X$.

A useful tool when proving statements about convergence in probability is Markov's inequality.

Theorem A.4.1 (Markov's inequality). *Let X be an \mathbb{R} -valued random variable. Then for any $0 < p < \infty$,*

$$\mathbb{P}(|X| \geq t) \leq \frac{1}{t^p} \int_{|X| \geq t} |X|^p d\mathbb{P}. \quad (\text{A.43})$$

For a proof of Markov's inequality see Ref.[118], which proves the general Chebyshev's inequality. Markov's inequality is a direct consequence of the general Chebyshev's inequality. The classical Chebyshev's inequality is a direct consequence of Markov's inequality.

For any $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$, such that $\mu = \mathbf{E}[Y]$ and $\sigma = \mathbf{E}[(Y - \mu)^2]$, and any $t > 0$,

by letting $p = 2$ in Markov's inequality,

$$\mathbb{P}(|Y - \mu| \geq t\sigma) = \mathbb{P}\left(\left|\frac{Y - \mu}{t}\right| \geq \sigma\right) \quad (\text{A.44})$$

$$\leq \frac{1}{t^2} \int_{|(Y - \mu)/t| \geq t} (Y - \mu)^2 d\mathbb{P} \quad (\text{A.45})$$

$$= \frac{\text{Var}[Y]}{t^2}, \quad (\text{A.46})$$

which is the classical form of Chebyshev's inequality.

Definition A.4.2 (Almost sure convergence). A sequence of random variables, $\{X_n\}_{n \in \mathbb{N}}$ converges almost surely to X if

$$\mathbb{P}\left(\lim_{n \rightarrow \infty} X_n = X\right) = 1. \quad (\text{A.47})$$

In other words,

$$\mathbb{P}(\{\omega \in \Omega : \lim_{n \rightarrow \infty} X_n(\omega) \neq X(\omega)\}) = 0. \quad (\text{A.48})$$

Lemma A.4.2 (Borel Cantelli lemmas). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\{A_n\}_{n \in \mathbb{N}}$ be a sequence in \mathcal{F} .

1. If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$ then the probability that infinitely many A_n occur is 0.

That is

$$\mathbb{P}\left(\omega \in \Omega : \omega \in \bigcap_{n=1}^{\infty} \bigcup_{k \geq n} A_k\right) = 0. \quad (\text{A.49})$$

2. If the A_n are mutual independent and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ then the probability

that infinitely many A_n occur is 1. That is,

$$\mathbb{P} \left(\omega \in \Omega : \omega \in \bigcap_{n=1}^{\infty} \bigcup_{k \geq n}^{\infty} A_k \right) = 1. \quad (\text{A.50})$$

Other types of convergence that are not used in this thesis include convergence in mean and convergence in distribution.

A.5 Jensen's Inequality

Theorem A.5.1 (Jensen's inequality). *Let $X \in \mathcal{L}^1(\Omega, \mathcal{F}, \mathbb{P})$, let S be a interval in \mathbb{R} (possibly infinite) that contains $X(\Omega)$, and let $\phi : S \rightarrow \mathbb{R}$ be a convex function. Then*

$$\mathbf{E}[\phi(X)] \geq \phi(\mathbf{E}[X]). \quad (\text{A.51})$$

Furthermore, if X is not constant a.s. and if ϕ is strictly convex then

$$\mathbf{E}[\phi(X)] > \phi(\mathbf{E}[X]). \quad (\text{A.52})$$

Proof. The proof is included because the proof of the strict inequality is difficult to find.

Because ϕ is convex, there is a linear function, $l : S \rightarrow \mathbb{R}$ such that

1. $l(x) \leq \phi(x) \quad \forall x \in S$
2. $l(\mathbf{E}[X]) = \phi(\mathbf{E}[X]),$

where it is known that $\mathbf{E}[X]$ is in the domain of both l and ϕ , which is S , because $X(\Omega) \subset S$ and S is convex.

Then

$$\phi(\mathbf{E}[X]) = l(\mathbf{E}[X]) \tag{A.53}$$

$$= \mathbf{E}[l(X)] \tag{A.54}$$

$$\leq \mathbf{E}[\phi(X)], \tag{A.55}$$

where the last inequality uses the first condition defining $l(x)$.

It is clear that strict inequality cannot hold if X is a constant, that is if $X = \mathbf{E}[X]$ a.s. If $X \neq \mathbf{E}[X]$ a.s. then let $A_{<} = \{x \in S : x < \mathbf{E}[X]\}$, let $A_{=} = \{x \in S : x = \mathbf{E}[x]\}$, let $A_{>} = \{x \in S : x > \mathbf{E}[x]\}$ and note that by the assumption $\mathbb{P}(X \in A_{<}) > 0$ and $\mathbb{P}(A_{>}) > 0$.

If ϕ is strictly convex, then, $l(x) > \phi(x)$ on $A_{<}$ and $A_{>}$ and $l(x) = \phi(x)$ on $A_{=}$, but since the push-forward measure of X on $A_{<}$ and $A_{>}$ is positive, it holds that

$$\mathbf{E}[l(X)] > \mathbf{E}[\phi(X)], \tag{A.56}$$

which strengthens inequality (A.55). □

A.6 The Radon-Nikodym theorem

Theorem A.6.1 (Radon-Nikodym theorem). *Let μ and ν be σ -finite measures on a measurable space (Ψ, \mathcal{A}) such that*

$$\mu \ll \nu. \tag{A.57}$$

Then there is an \mathcal{A} -measurable function $g : \Psi \rightarrow [0, \infty]$, such that for all $A \in \mathcal{A}$,

$$\mu(A) = \int_A g \, d\nu. \quad (\text{A.58})$$

The function g is called the Radon-Nikodym derivative of μ with respect to ν and is denoted $\frac{d\mu}{d\nu}$.

Furthermore, any two Radon-Nikodym derivatives of μ with respect to ν are equivalent up to a set of measure 0.

The fundamental importance of the Radon-Nikodym stems from the fact that σ -algebras and measures on σ -algebras can be very difficult to work with due to large cardinality. For instance, the Lebesgue measurable sets have the same cardinality as $\mathcal{P}(\mathbb{R})$. It is difficult to study arbitrary measures on σ -algebras with such high cardinality because it could be impossible to approximate them by countable sequences of measures. The Radon-Nikodym derivative, if it exists, contains all of the information that the measure has, but, since it is a real-valued function, it is possible to approximate the function by countable sequences of simpler functions.

The uniqueness follows from considering the function $g = g_1 - g_2$, where g_1 and g_2 are Radon-Nikodym derivatives for μ with respect to ν .

If it has been shown that the Radon-Nikodym theorem holds for finite positive measures then it is fairly straightforward to extend that result to σ -finite measures and signed measures. In particular, a σ -finite space can be written as the disjoint union of a countable sequence of measurable sets, $\{B_i\}_{i=1}^\infty$ such that $\mu(B_i) < \infty$ and $\nu(B_i) < \infty$. By the Radon-Nikodym theorem for finite measures, for each n there is a measurable $g_n : B_n \rightarrow [0, \infty]$ such that for all $A \in \mathcal{A}$,

$$\mu(A) = \int_{A \cap B_n} g_n \, d\nu. \quad (\text{A.59})$$

The function $g : \Psi \rightarrow [0, \infty]$ be defined by $g|_{B_n} = g_n$ is the Radon-Nikodym derivative of μ with respect to ν .

If μ is a signed σ -finite measure and ν is a positive σ -finite measure then there is a Jordan decomposition [104] of μ into a positive and a negative part such that

$$\mu = \mu^+ - \mu^-. \quad (\text{A.60})$$

By the result for positive σ -finite measures there are corresponding Radon-Nikodym derivatives h^+ and h^- , and the difference

$$h = h^+ - h^- \quad (\text{A.61})$$

is the required Radon-Nikodym derivative.

The following proof of the theorem for finite measures follows the outline provided in Exercise 59 of Section 18.4 in Real Analysis by Royden and Fitzpatrick [104]. The proof employs the framework of functional analysis by representing the integral as a linear functional on a Hilbert space of measurable functions. The integral is then represented by a measurable function using the Riesz representation theorem.

Proposition A.6.2. *Let μ and ν be finite measures on the measurable space (Ψ, \mathcal{A}) . Then there is an \mathcal{A} -measurable function $g : \Psi \rightarrow [0, \infty]$ such that for all $A \in \mathcal{A}$,*

$$\mu(A) = \int_A g \, d\nu \quad (\text{A.62})$$

Proof. Let $\pi = \nu + \mu$ so that $\nu \ll \pi$ and $\mu \ll \pi$. For any $f \in \mathcal{L}^2(\Psi, \pi)$ define

$$T(f) = \int_{\Psi} f \, d\nu. \quad (\text{A.63})$$

Then T is a bounded linear functional on $\mathcal{L}^2(\Psi, \pi)$. The boundedness follows from the finiteness of ν . In particular, if $f \in \mathcal{L}^2(\Psi, \pi)$, then since

$$\int f^2 d\nu + \int f^2 d\mu = \int f^2 d\pi < \infty \quad (\text{A.64})$$

and $\int f^2, d\mu > 0$, it follows that $\int f^2 d\nu < \int f^2 d\pi$. Furthermore,

$$(T(f))^2 = \left(\int_{\Psi} f \cdot 1 d\nu \right)^2 \quad (\text{A.65})$$

$$\leq \left(\int_{\Psi} f^2 d\nu \right) \left(\int_{\Psi} d\nu \right) \quad (\text{A.66})$$

$$= \nu(\Psi) \int_{\Psi} f^2 d\nu \quad (\text{A.67})$$

$$\leq \nu(\Psi) \int_{\Psi} f^2 d\pi, \quad (\text{A.68})$$

so that

$$\|T\| = \sup \{ |T(f)| : \|f\|_2 = 1 \} \quad (\text{A.69})$$

$$\leq \sqrt{\nu(\Psi)} \quad (\text{A.70})$$

$$< \infty. \quad (\text{A.71})$$

Since T is a bounded linear functional, by the Riesz representation theorem, there is an $h \in \mathcal{L}^2(\Psi, \pi)$ such that

$$T(f) = \int_{\Psi} f \cdot h d\pi \quad (\text{A.72})$$

for all $f \in \mathcal{L}^2(\Psi, \pi)$.

The required function g can be derived from h . Substitution the definitions for T

and π into Equation A.72 yields

$$\int_{\Psi} f d\nu = \int_{\Psi} f \cdot h d\nu + \int_{\Psi} f \cdot h d\mu. \quad (\text{A.73})$$

Letting $f = \mathcal{X}_A$ yields

$$\nu(A) = \int_A h d\nu + \int_A h d\mu = \int_A h d\pi. \quad (\text{A.74})$$

Substituting $A = \{\psi \in \Psi : h \leq 0\}$ reveals that $h > 0$ μ and ν -almost everywhere.

Therefore, $1/h$ is ν , μ , and π -measurable, and by Eq. A.72,

$$\int_A \frac{1}{h} d\nu = \int_A \frac{1}{h} \cdot h d\pi = \pi(A). \quad (\text{A.75})$$

Therefore,

$$\mu(A) = \pi(A) - \nu(A) \quad (\text{A.76})$$

$$= \int_A \frac{1}{h} - 1 d\nu. \quad (\text{A.77})$$

Therefore, the required function is

$$g = \frac{1}{h} - 1. \quad (\text{A.78})$$

□

Bibliography

- [1] Sophie Achard and Ed Bullmore. “Efficiency and cost of economical brain functional networks”. In: *PLoS computational biology* 3.2 (2007), e17.
- [2] Sophie Achard, Raymond Salvador, Brandon Whitcer, John Suckling, and ED Bullmore. “A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs”. In: *Journal of Neuroscience* 26.1 (2006), pp. 63–72.
- [3] Nicola Ancona, Daniele Marinazzo, and Sebastiano Stramaglia. “Radial basis function approach to nonlinear Granger causality of time series”. In: *Physical Review E* 70.5 (2004), p. 056221.
- [4] Mihael Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jörg Sander. “OPTICS: ordering points to identify the clustering structure”. In: *ACM Sigmod record*. Vol. 28. 2. ACM. 1999, pp. 49–60.
- [5] M Ballerini et al. “Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study”. In: *Proceedings of the National Academy of Sciences of the United States of America* 105.4 (2008), pp. 1232–1237.
- [6] Ernest Barreto, Brian Hunt, Edward Ott, and Paul So. “Synchronization in networks of networks: The onset of coherent collective behavior in systems

- of interacting populations of heterogeneous oscillators”. In: *Phys. Rev. E* 77 (2008), p. 036107.
- [7] Adam B Barrett, Lionel Barnett, and Anil K Seth. “Multivariate Granger causality and generalized variance”. In: *Physical Review E* 81.4 (2010), p. 041907.
- [8] William Beckner. “Inequalities in Fourier analysis”. In: *Annals of Mathematics* (1975), pp. 159–182.
- [9] Jan Beirlant, Edward J Dudewicz, László Györfi, and Edward C Van der Meulen. “Nonparametric entropy estimation: An overview”. In: *International Journal of Mathematical and Statistical Sciences* 6.1 (1997), pp. 17–39.
- [10] Gabriele Bellucci, Sergey Chernyak, Morris Hoffman, Gopikrishna Deshpande, Olga Dal Monte, Kristine M Knutson, Jordan Grafman, and Frank Krueger. “Effective connectivity of brain regions underlying third-party punishment: functional MRI and Granger causality evidence”. In: *Social neuroscience* 12.2 (2017), pp. 124–134.
- [11] Jon Louis Bentley. “Multidimensional binary search trees used for associative searching”. In: *Communications of the ACM* 18.9 (1975), pp. 509–517.
- [12] Iwo Bialynicki-Birula and Łukasz Rudnicki. “Entropic uncertainty relations in quantum physics”. In: *Statistical Complexity*. Springer, 2011, pp. 1–34.
- [13] Susan Blackmore. *The meme machine*. Vol. 25. Oxford Paperbacks, 2000.
- [14] Adam Bobrowski. *Functional analysis for probability and stochastic processes: an introduction*. Cambridge University Press, 2005.
- [15] N. W. F. Bode, A. J. Wood, and D. W. Franks. “Social networks and models for collective motion in animals”. In: *Behav. Ecol. Sociobiol.* 65 (2011), pp. 117–130.

- [16] N. W. F. Bode, A. J. Wood, and D. W. Franks. “The impact of social networks on animal collective motion”. In: *Anim. Behav.* 82 (2011), pp. 29–38.
- [17] Erik M Bollt. “Synchronization as a process of sharing and transferring information”. In: *International Journal of Bifurcation and Chaos* 22.11 (2012), p. 1250261.
- [18] Erik M Bollt and Naratip Santitissadeekorn. *Applied and Computational Measurable Dynamics*. SIAM, 2013.
- [19] Steven L Bressler and Anil K Seth. “Wiener–Granger causality: a well established methodology”. In: *Neuroimage* 58.2 (2011), pp. 323–329.
- [20] Carlo Cafaro, Warren M Lord, Jie Sun, and Erik M Bollt. “Causation entropy from symbolic representations of dynamical systems”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 25.4 (2015), p. 043106.
- [21] Rafael S Calsaverini and Renato Vicente. “An information-theoretic approach to statistical dependence: Copula information”. In: *EPL (Europhysics Letters)* 88.6 (2009), p. 68003.
- [22] Mario Chavez, Jacques Martinerie, and Michel Le Van Quyen. “Statistical assessment of nonlinear causality: application to epileptic EEG signals”. In: *Journal of Neuroscience Methods* 124 (2003), pp. 113–128.
- [23] Ronald R Coifman and Stéphane Lafon. “Diffusion maps”. In: *Applied and computational harmonic analysis* 21.1 (2006), pp. 5–30.
- [24] I. D. Couzin, J. Krause, R. James, G. D. Ruxton, and N. R. Franks. “Collective memory and spatial sorting in animal groups”. In: *J. Theor. Biol.* 218 (2002), pp. 1–11.

- [25] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2012.
- [26] Thomas M Cover and Zhen Zhang. “On the maximum entropy of the sum of two dependent random variables”. In: *IEEE Transactions on Information Theory* 40.4 (1994), pp. 1244–1246.
- [27] Imre Csiszár. “Axiomatic characterizations of information measures”. In: *Entropy* 10.3 (2008), pp. 261–273.
- [28] Georges A Darbellay and Igor Vajda. “Estimation of the information by an adaptive partitioning of the observation space”. In: *IEEE Transactions on Information Theory* 45.4 (1999), pp. 1315–1321.
- [29] Thomas S Deisboeck and Iain D Couzin. “Collective behavior in cancer cell populations”. In: *BioEssays* 31 (2009), pp. 190–197.
- [30] A Dembo. *Probability Theory: STAT310/MATH230*. University Lecture. 2016. URL: <http://statweb.stanford.edu/~adembo/stat-310b/lnotes.pdf>.
- [31] Amir Dembo, Thomas M Cover, and Joy A Thomas. “Information theoretic inequalities”. In: *IEEE Transactions on Information Theory* 37.6 (1991), pp. 1501–1518.
- [32] Yu G Dmitriev and FP Tarasenko. “On the estimation of functionals of the probability density and its derivatives”. In: *Theory of Probability & Its Applications* 18.3 (1974), pp. 628–633.
- [33] J. A. Downes. “The swarming and mating flight of *Diptera*”. In: *Annu. Rev. Entomol.* 14 (1969), pp. 271–298.

- [34] Victor M Eguiluz, Dante R Chialvo, Guillermo A Cecchi, Marwan Baliki, and A Vania Apkarian. “Scale-free brain functional networks”. In: *Physical review letters* 94.1 (2005), p. 018102.
- [35] Joseph Emerson, Yaakov S Weinstein, Marcos Saraceno, Seth Lloyd, and David G Cory. “Pseudo-random unitary operators for quantum information processing”. In: *science* 302.5653 (2003), pp. 2098–2100.
- [36] Luca Ferrarini, Ilya M Veer, Evelinda Baerends, Marie-José van Tol, Remco J Renken, Nic JA van der Wee, Dirk Veltman, Andre Aleman, Frans G Zitman, Brenda WJH Penninx, et al. “Hierarchical functional modularity in the resting-state human brain”. In: *Human brain mapping* 30.7 (2009), pp. 2220–2231.
- [37] Yoav Freund, Sanjoy Dasgupta, Mayank Kabra, and Nakul Verma. “Learning the structure of manifolds using random projections”. In: *Advances in Neural Information Processing Systems*. 2008, pp. 473–480.
- [38] Jerome H Friedman, Werner Stuetzle, and Anne Schroeder. “Projection pursuit density estimation”. In: *Journal of the American Statistical Association* 79.387 (1984), pp. 599–608.
- [39] K.J. Friston, A. Mechelli, R. Turner, and C.J. Price. “Nonlinear Responses in fMRI: The Balloon Model, Volterra Kernels, and Other Hemodynamics”. In: *NeuroImage* 12.4 (2000), pp. 466–477.
- [40] Shuyang Gao, Greg Ver Steeg, and Aram Galstyan. “Efficient estimation of mutual information for strongly dependent variables”. In: *Artificial Intelligence and Statistics*. 2015, pp. 277–286.
- [41] Weihao Gao, Sewoong Oh, and Pramod Viswanath. “Demystifying fixed k-nearest neighbor information estimators”. In: *IEEE Transactions on Information Theory* (2018).

- [42] Hamid Ghourchian, Amin Gohari, and Arash Amini. “Existence and continuity of differential entropy for a class of distributions”. In: *IEEE Communications Letters* 21.7 (2017), pp. 1469–1472.
- [43] Irene Giardina. “Collective behavior in animal groups: theoretical models and empirical studies”. In: *Human Frontier Science Program Journal* 2.4 (2008), pp. 205–219.
- [44] Maria Teresa Giraudo, Laura Sacerdote, and Roberta Sirovich. “Non-parametric estimation of mutual information through the entropy of the linkage”. In: *Entropy* 15.12 (2013), pp. 5154–5177.
- [45] Gene H Golub and Charles F Van Loan. *Matrix computations*. Vol. 3. JHU Press, 2012.
- [46] Peter beim Graben, Kristin K Sellers, Flavio Fröhlich, and Axel Hutt. “Optimal estimation of recurrence structures from time series”. In: *EPL (Europhysics Letters)* 114.3 (2016), p. 38003.
- [47] Clive WJ Granger. “Investigating causal relations by econometric models and cross-spectral methods”. In: *Econometrica: Journal of the Econometric Society* (1969), pp. 424–438.
- [48] Clive WJ Granger. “Testing for causality: a personal viewpoint”. In: *Journal of Economic Dynamics and control* 2 (1980), pp. 329–352.
- [49] Thomas Gregor, Koichi Fujimoto, Noritaka Masaki, and Satoshi Sawai. “The Onset of Collective Behavior in Social Amoebae”. In: *Science* 328.5981 (2010), pp. 1021–1025.
- [50] Alexander Grigoryan. *Measure theory and probability*. University Lecture. 2007. URL: <https://www.math.uni-bielefeld.de/~grigor/mwlect.pdf>.

- [51] Paul R Halmos. *Measure theory*. Vol. 18. Springer, 2013.
- [52] Frank Hampel. “Robust statistics: A brief introduction and overview”. In: *First International Symposium on Robust Statistics and Fuzzy Techniques in Geodesy and GIS*. A. Carosio, H. Kutterer (editors), Swiss Federal Institute of Technology Zurich (ETH), Institute of Geodesy and Photogrammetry, IGP-Bericht. 295. 2001, pp. 13–17.
- [53] Frank R Hampel. “The influence curve and its role in robust estimation”. In: *Journal of the american statistical association* 69.346 (1974), pp. 383–393.
- [54] Mario Hentschel, Daniel Dregely, Ralf Vogelgesang, Harald Giessen, and Na Liu. “Plasmonic Oligomers: The Role of Individual Particles in Collective Behavior”. In: *ACS Nano* 5.3 (2011), pp. 2042–2050.
- [55] James E Herbert-Read, Andrea Perna, Richard P Mann, Timothy M Schaerf, David JT Sumpter, and Ashley JW Ward. “Inferring the rules of interaction of shoaling fish”. In: *Proceedings of the National Academy of Sciences* 108.46 (2011), pp. 18726–18731.
- [56] Suzana Herculano-Houzel. “The remarkable, yet not extraordinary, human brain as a scaled-up primate brain and its associated cost”. In: *Proceedings of the National Academy of Sciences* 109.Supplement 1 (2012), pp. 10661–10668.
- [57] Peter Hoff. *Invariant inference*. University Lecture. 2013. URL: <https://www.stat.washington.edu/~pdhoff/courses/581/LectureNotes/equivariance.pdf>.
- [58] Peter J Huber. “The 1972 wald lecture robust statistics: A review”. In: *The Annals of Mathematical Statistics* (1972), pp. 1041–1067.

- [59] Ryan G James, Nix Barnett, and James P Crutchfield. “Information flows? A critique of transfer entropies”. In: *Physical review letters* 116.23 (2016), p. 238701.
- [60] Harry Joe. “Estimation of entropy and other functionals of a multivariate density”. In: *Annals of the Institute of Statistical Mathematics* 41.4 (1989), pp. 683–697.
- [61] Yael Katz, Kolbjørn Tunstrøm, Christos C Ioannou, Cristián Huepe, and Iain D Couzin. “Inferring the structure and dynamics of interactions in schooling fish”. In: *Proceedings of the National Academy of Sciences* 108.46 (2011), pp. 18720–18725.
- [62] D. H. Kelley and N. T. Ouellette. “Emergent dynamics of laboratory insect swarms”. In: *Sci. Rep.* 3 (2013), p. 1073.
- [63] A Ya Khinchin. *Mathematical foundations of information theory*. Courier Corporation, 2013.
- [64] Pileun Kim, Jonathan Rogers, Jie Sun, and Erik Bollt. “Causation entropy identifies sparsity structure for parameter estimation of dynamic systems”. In: *Journal of Computational and Nonlinear Dynamics* 12.1 (2017), p. 011008.
- [65] Richard Kleeman. “Measuring dynamical prediction utility using relative entropy”. In: *Journal of the atmospheric sciences* 59.13 (2002), pp. 2057–2072.
- [66] Jon M Kleinberg. “Navigation in a small world”. In: *Nature* 406.6798 (2000), p. 845.
- [67] LF Kozachenko and Nikolai N Leonenko. “Sample estimate of the entropy of a random vector”. In: *Problemy Peredachi Informatsii* 23.2 (1987), pp. 9–16.

- [68] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. “Estimating mutual information”. In: *Physical review E* 69.6 (2004), p. 066138.
- [69] Thomas Kreuz, Florian Mormann, Ralph G Andrzejak, Alexander Kraskov, Klaus Lehnertz, and Peter Grassberger. “Measuring synchronization in coupled model systems: A comparison of different approaches”. In: *Physica D: Nonlinear Phenomena* 225.1 (2007), pp. 29–42.
- [70] Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.
- [71] X San Liang and Richard Kleeman. “Information transfer between dynamical system components”. In: *Physical review letters* 95.24 (2005), pp. 244101–244101.
- [72] Warren M Lord, Jie Sun, and Erik M Bollt. “Geometric k-nearest neighbor estimation of entropy and mutual information”. In: *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28.3 (2018), p. 033114.
- [73] Warren M Lord, Jie Sun, Nicholas T Ouellette, and Erik M Bollt. “Inference of causal information flow in collective animal behavior”. In: *IEEE Transactions on Molecular, Biological and Multi-Scale Communications* 2.1 (2016), pp. 107–116.
- [74] Ryan Lukeman, Yue-Xian Li, and Leah Edelstein-Keshet. “Inferring individual rules from collective behavior”. In: *Proceedings of the National Academy of Sciences* 107.28 (2010), pp. 12576–12580.
- [75] Y P Mack and M Rosenblatt. “Multivariate k-nearest neighbor density estimates”. In: *Journal of Multivariate Analysis* 9.1 (1979), pp. 1–15. ISSN: 0047-259X.

- [76] Mokshay Madiman and Ioannis Kontoyiannis. “The entropies of the sum and the difference of two IID random variables are not too different”. In: *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*. IEEE. 2010, pp. 1369–1372.
- [77] Prasanta Chandra Mahalanobis. “On the generalised distance in statistics”. In: *Proceedings of the National Institute of Sciences of India, 1936* (1936), pp. 49–55.
- [78] David Meunier, Sophie Achard, Alexa Morcom, and Ed Bullmore. “Age-related changes in modular organization of human brain functional networks”. In: *Neuroimage* 44.3 (2009), pp. 715–723.
- [79] Abdelkader Mokkadem. “Estimation of the entropy and information of absolutely continuous random variables”. In: *IEEE Transactions on Information Theory* 35.1 (1989), pp. 193–196.
- [80] Mehdi Moussaid, Simon Garnier, Guy Theraulaz, and Dirk Helbing. “Collective Information Processing and Pattern Formation in Swarms, Flocks, and Crowds”. In: *Topics in Cognitive Science* 1 (2009), pp. 469–497.
- [81] Mehdi Moussaïda, Dirk Helbing, and Guy Theraulaz. “How simple rules determine pedestrian behavior and crowd disasters”. In: *Proceedings of the National Academy of Sciences of the United States of America* 108.17 (2011), pp. 6884–6888.
- [82] Mark Newman. “The Structure and Function of Complex Networks”. In: *SIAM Review* 45.2 (2003), pp. 167–256.
- [83] Mark Newman. *Networks: an introduction*. Oxford university press, 2010.
- [84] R. Ni and N. T. Ouellette. “Velocity correlations in laboratory insect swarms”. In: *Eur. Phys. J. Special Topics* 224 (2015), pp. 3271–3277.

- [85] R. Ni and N. T. Ouellette. “On the tensile strength of insect swarms”. In: *submitted* (2016).
- [86] R. Ni, J. G. Puckett, E. R. Dufresne, and N. T. Ouellette. “Intrinsic fluctuations and driven response of insect swarms”. In: *Physical Review Letters* 115 (2015), p. 118104.
- [87] N Orange and N Abaid. “A transfer entropy analysis of leader-follower interactions in flying bats”. In: *The European Physical Journal Special Topics* 224.17-18 (2015), pp. 3279–3293.
- [88] Edward Ott. *Chaos in dynamical systems*. Cambridge university press, 2002.
- [89] N. T. Ouellette. “Empirical questions for collective-behaviour modelling”. In: *Pramana-J. Phys.* 84 (2015), pp. 353–363.
- [90] N. T. Ouellette, H. Xu, and E. Bodenschatz. “A quantitative study of three-dimensional Lagrangian particle tracking algorithms”. In: *Exp. Fluids* 40 (2006), pp. 301–313.
- [91] J. K. Parrish and L. Edelstein-Keshet. “Complexity, pattern, and evolutionary trade-offs in animal aggregation”. In: *Science* 284 (1999), pp. 99–101.
- [92] Brian L Partridge. “The structure and function of fish schools”. In: *Scientific american* 246.6 (1982), pp. 114–123.
- [93] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2000.
- [94] Yury Polyanskiy and Yihong Wu. “Wasserstein continuity of entropy and outer bounds for interference channels”. In: *IEEE Transactions on Information Theory* 62.7 (2016), pp. 3992–4002.

- [95] Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, et al. “Functional network organization of the human brain”. In: *Neuron* 72.4 (2011), pp. 665–678.
- [96] J. G. Puckett, D. H. Kelley, and N. T. Ouellette. “Searching for effective forces in laboratory insect swarms”. In: *Sci. Rep.* 4 (2014), p. 4766.
- [97] J. G. Puckett, R. Ni, and N. T. Ouellette. “Time-frequency analysis reveals pairwise interactions in insect swarms”. In: *Physical Review Letters* 114 (2015), p. 258103.
- [98] J. G. Puckett and N. T. Ouellette. “Determining asymptotically large population sizes in insect swarms”. In: *J. R. Soc. Interface* 11 (2014), p. 20140710.
- [99] Svetlozar T Rachev, Lev Klebanov, Stoyan V Stoyanov, and Frank Fabozzi. *The methods of distances in the theory of probability and statistics*. Springer Science & Business Media, 2013.
- [100] BLS Prakasa Rao. *Nonparametric functional estimation*. Academic press, 2014.
- [101] Bhargava Ravoori, Adam B Cohen, Jie Sun, Adilson E Motter, Thomas E Murphy, and Rajarshi Roy. “Robustness of Optimal Synchronization in Real Networks”. In: *Physical Review Letters* 107 (2011), p. 034102.
- [102] C. W. Reynolds. “Flocks, herds, and schools: A distributed behavioral model”. In: *SIGGRAPH Comput. Graph.* 21 (1987), pp. 25–34.
- [103] Monica D Rosenberg, Emily S Finn, Dustin Scheinost, Xenophon Papademetris, Xilin Shen, R Todd Constable, and Marvin M Chun. “A neuromarker of sustained attention from whole-brain functional connectivity”. In: *Nature neuroscience* 19.1 (2016), p. 165.

- [104] Halsey Lawrence Royden and Patrick Fitzpatrick. *Real analysis*. Vol. 2. Macmillan New York, 1968.
- [105] Raymond Salvador, John Suckling, Martin R Coleman, John D Pickard, David Menon, and ED Bullmore. “Neurophysiological architecture of functional magnetic resonance images of human brain”. In: *Cerebral cortex* 15.9 (2005), pp. 1332–1342.
- [106] X San Liang and Richard Kleeman. “A rigorous formalism of information transfer between dynamical system components. I. Discrete mapping”. In: *Physica D: Nonlinear Phenomena* 231.1 (2007), pp. 1–9.
- [107] X San Liang and Richard Kleeman. “A rigorous formalism of information transfer between dynamical system components. II. Continuous flow”. In: *Physica D: Nonlinear Phenomena* 227.2 (2007), pp. 173–182.
- [108] Thomas Schreiber. “Measuring information transfer”. In: *Physical review letters* 85.2 (2000), p. 461.
- [109] CR Shalizi and A Kontorovich. *Almost None of the Theory of Stochastic Processes*. 2010. URL: <http://www.stat.cmu.edu/~cshalizi/almost-none/>.
- [110] C. E. Shannon. “A mathematical theory of communication”. In: *Bell System Tech. J.* 27 (1948), pp. 379–423, 623–656.
- [111] J Shao. *Mathematical Statistics, Springer Texts in Statistics*. 2003.
- [112] Harshinder Singh, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk. “Nearest neighbor estimates of entropy”. In: *American journal of mathematical and management sciences* 23.3-4 (2003), pp. 301–321.
- [113] Per Sebastian Skardal, Dane Taylor, and Jie Sun. “Optimal synchronization of complex networks”. In: *Physical Review Letters* 113 (2014), p. 144101.

- [114] Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. “Network modelling methods for FMRI”. In: *Neuroimage* 54.2 (2011), pp. 875–891.
- [115] Andrey Sokolov, Igor S Aranson, John O Kessler, and Raymond E Goldstein. “Concentration Dependence of the Collective Dynamics of Swimming Bacteria”. In: *Physical Review Letters* 98 (2007), p. 158102.
- [116] Lee Spector, Jon Klein, Chris Perry, and Mark Feinstein. “Emergence of Collective Behavior in Evolving Populations of Flying Agents”. In: *Genetic Programming and Evolvable Machines* 6 (2005), pp. 111–125.
- [117] Kumar Sricharan, Raviv Raich, and Alfred O Hero III. “Estimation of Non-linear Functionals of Densities With Confidence”. In: *IEEE Transactions on Information Theory* 58.7 (2012), pp. 4135–4159.
- [118] George Stepaniants. *The Lebesgue Integral, Chebyshev’s Inequality, and the Weierstrass Approximation Theorem*. Webpage. 2017. URL: https://sites.math.washington.edu/~morrow/336_17/papers17/george.pdf.
- [119] Dan Stowell and Mark D Plumbley. “Fast Multidimensional Entropy Estimation by k -d Partitioning”. In: *IEEE Signal Processing Letters* 16.6 (2009), pp. 537–540.
- [120] David J T Sumpter. *Collective Animal Behavior*. Princeton University Press, 2010.
- [121] Jie Sun and Erik M Bollt. “Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings”. In: *Physica D: Nonlinear Phenomena* 267 (2014), pp. 49–57.

- [122] Jie Sun, Erik M. Bollt, Mason A Porter, and Marian S Dawkins. “A Mathematical Model for the Dynamics and Synchronization of Cows”. In: *Physica D* 240 (2011), pp. 1497–1509.
- [123] Jie Sun, Carlo Cafaro, and Erik M Bollt. “Identifying the coupling structure in complex systems through the optimal causation entropy principle”. In: *Entropy* 16.6 (2014), pp. 3416–3433.
- [124] Jie Sun, Dane Taylor, and Erik M Bollt. “Causal network inference by optimal causation entropy”. In: *SIAM Journal on Applied Dynamical Systems* 14.1 (2015), pp. 73–106.
- [125] M. Tennenbaum, Z. Liu, D. Hu, and A. Fernandez-Nieves. “Mechanics of fire ant aggregations”. In: *Nat. Mater.* 15 (2016), pp. 54–59.
- [126] Takenori Tomaru, Hisashi Murakami, Takayuki Niizato, Yuta Nishiyama, Kohei Sonoda, Toru Moriyama, and Yukio-Pegio Gunji. “Information transfer in a swarm of soldier crabs”. In: *Artificial Life and Robotics* 21.2 (2016), pp. 177–180.
- [127] J. Toner, Y. Tu, and S. Ramaswamy. “Hydrodynamics and phases of flocks”. In: *Ann. Phys.* 318 (2005), pp. 170–244.
- [128] R. Y. Tsai. “A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses”. In: *IEEE J. Robot. Autom.* RA-3 (1987), pp. 323–344.
- [129] John W Tukey. “A survey of sampling from contaminated distributions”. In: *Contributions to probability and statistics* (1960), pp. 448–485.
- [130] Martin Vejmelka and Milan Paluš. “Inferring the directionality of coupling with conditional mutual information”. In: *Physical Review E* 77.2 (2008), p. 026214.

- [131] PF Verdes. “Assessing causality from multivariate time series”. In: *Physical Review E* 72.2 (2005), p. 026222.
- [132] Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. “Transfer entropy—a model-free measure of effective connectivity for the neurosciences”. In: *J. Comput. Neurosci.* 30 (2011), pp. 45–67.
- [133] Raul Vicente, Michael Wibral, Michael Lindner, and Gordon Pipa. “Transfer entropy—a model-free measure of effective connectivity for the neurosciences”. In: *Journal of computational neuroscience* 30.1 (2011), pp. 45–67.
- [134] T. Vicsek, A. Czirók, E. Ben-Jacob, I. Cohen, and O. Shochet. “Novel type of phase transition in a system of self-driven particles”. In: *Physical Review Letters* 75 (1995), pp. 1226–1229.
- [135] Duncan J Watts and Steven H Strogatz. “Collective dynamics of ‘small-world’ networks”. In: *Nature* 393.6684 (1998), p. 440.
- [136] Norbert Wiener. “The theory of prediction”. In: *Modern mathematics for engineers* (1956).
- [137] Paul L Williams and Randall D Beer. “Nonnegative decomposition of multivariate information”. In: *arXiv preprint arXiv:1004.2515* (2010).
- [138] Paul L Williams and Randall D Beer. “Generalized measures of information transfer”. In: *arXiv preprint arXiv:1102.1507* (2011).
- [139] Simon Wing, Jay R Johnson, and Angelos Vourlidas. “Information Theoretic Approach to Discovering Causalities in the Solar Cycle”. In: *The Astrophysical Journal* 854.2 (2018), p. 85.

- [140] PR Wozniak and A Kruszewski. “On Estimating Non-Uniform Density Distributions Using N Nearest Neighbors”. In: *Acta Astronomica* 62 (2012), pp. 409–417.
- [141] Haitao Xu. “Tracking Lagrangian trajectories in position–velocity space”. In: *Measurement Science and Technology* 19.7 (2008), p. 075105.
- [142] Christian A Yatesa, Radek Erban, Carlos Escudero, Iain D Couzind, Jerome Buhl, Ioannis G Kevrekidis, Philip K Maini, and David J T Sumpter. “Inherent noise can facilitate coherence in collective swarm motion”. In: *Proceedings of the National Academy of Sciences of the United States of America* 106.14 (2009), pp. 5464–5469.
- [143] H P Zhang, Avraham Be’er, E-L Florin, and Harry L Swinney. “Collective motion and density fluctuations in bacterial colonies”. In: *Proceedings of the National Academy of Sciences of the United States of America* 107.31 (2010), pp. 13626–13630.
- [144] Jie Zhu, Jean-Jacques Bellanger, Huazhong Shu, Chunfeng Yang, and Régine Le Bouquin Jeannès. “Bias reduction in the estimation of mutual information”. In: *Physical Review E* 90.5 (2014), p. 052714.
- [145] Gordan Žitković. University Lecture. 2010. URL: https://www.ma.utexas.edu/users/gordanz/notes/theory_of_probability_I.pdf.