

Estimation of Mutual Information

M.Sc. Thesis

by

Harsha Rahulkumar Jain



DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY INDORE
MAY 2026

Estimation of Mutual Information

A THESIS

*Submitted in partial fulfillment of the requirements for the award of
the degree of*

Master of Science

by

Harsha Rahulkumar Jain

(Roll No. 2403141004)

Under the guidance of

Dr. Mohd. Arshad



**DEPARTMENT OF MATHEMATICS
INDIAN INSTITUTE OF TECHNOLOGY INDORE
MAY 2026**

INDIAN INSTITUTE OF TECHNOLOGY INDORE

CANDIDATE'S DECLARATION

I hereby certify that the work which is being presented in the thesis entitled “**Estimation of Mutual Information**” in the partial fulfillment of the requirements for the award of the degree of **MASTER OF SCIENCE** and submitted in the **DEPARTMENT OF MATHEMATICS, INDIAN INSTITUTE OF TECHNOLOGY INDORE**, is an authentic record of my own work carried out during the time period from July 2025 to May 2026 under the supervision of **Dr. Mohd. Arshad**, Associate Professor, Department of Mathematics, IIT Indore. The matter presented in this thesis by me has not been submitted for the award of any other degree of this or any other institute.

Signature of the student with date

21/05/26

(**Harsha Rahulkumar Jain**)

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

21-05-2026

Signature of Thesis Supervisor with date

(**Dr. Mohd. Arshad**)

Harsha Rahulkumar Jain has successfully given her M.Sc. Oral Examination held on 13th May, 2026.

Signature of Supervisor of M.Sc. Thesis

Date: **21-05-2026**

Signature of Convener, DPGC

Date: **21/05/2026**

*Dedicated to my
Parents and Rahul*

“ Those times when you get up early and you work hard, those times when you stay up late and you work hard, those times when you don’t feel like working, you’re too tired, you don’t want to push yourself, but you do it anyway. That is actually the dream. That’s the dream. It’s not the destination, it’s the journey. And if you guys can understand that, then what you’ll see happen is you won’t accomplish your dreams, your dreams won’t come true; something greater will.”

– Kobe Bryant

Acknowledgements

Deepest gratitude is expressed to the supervisor, Dr Mohd. Arshad, for constant guidance, invaluable support, and insightful suggestions throughout the academic journey. His mentorship has been instrumental in shaping understanding of mathematics and research. His analytical approach to problem-solving and commitment to academic excellence have been a constant source of inspiration. This work would not have been possible without his encouragement and dedicated supervision. I sincerely thank the members of my committee, Prof. Md. Aquil Khan, Prof. Bapan Ghosh, Dr Bibekananda Maji, and Dr Charitha Cherugondi, for their valuable suggestions, constructive feedback, and continuous support throughout this work. I also acknowledge the support and academic encouragement from senior research scholars and colleagues in the department, whose discussions and suggestions contributed significantly to my learning and research development. Sincere appreciation is extended to the administrative and support staff for their dedicated assistance in ensuring a smooth academic experience. The library staff and the IT support team are also acknowledged for providing essential academic resources and technical assistance throughout the

research work. I gratefully acknowledge the Department of Science and Technology (DST), India, for providing access to the Bhaskaracharya Lab and its excellent facilities, which significantly contributed to the successful completion of this work. I owe my deepest gratitude to my family for their unconditional love, encouragement, patience, and unwavering faith in me. Their constant support has been my greatest source of strength and motivation throughout this journey. Finally, I express my heartfelt gratitude to everyone who has directly or indirectly contributed to my academic and personal growth. Their support, encouragement, and goodwill have played an important role in the successful completion of this work.

Abstract

Geometric partition entropy is a recently developed non-parametric approach for entropy estimation, which gives a different way of looking at the problem by taking into account the effect of informative outliers through geometric structures. However, the initial definition was limited only to one-dimensional state spaces. In this work, we considered a generalised version of geometric partition entropy for samples lying in a bounded region of an n -dimensional vector space. This generalized definition is based on the concept of a Voronoi diagram, where the entropy is computed using the measures of the corresponding geometric partitions. Although this approach is theoretically meaningful, constructing Voronoi diagrams becomes computationally expensive and often unreliable as the dimension increases.

To overcome this issue, approximation methods have been developed that speed up computation while maintaining the same accuracy. Among these, the partition intersection (π) approximation is particularly useful, as it allows direct estimation of marginal entropy and can also be used to build an efficient mutual information estimator.

Further, to handle high-dimensional data more effectively, a granular ball-based entropy estimator is proposed. Instead of relying on complicated geometric partitions, this method uses adaptive ball structures to represent the data. This method not only reduces computational cost but also makes the estimator more robust to outliers and scalable to higher dimensions.

Overall, this thesis contributes to develop a new estimator for mutual information by adding the granular ball computing theory.

Contents

Abstract	v
List of Figures	ix
1 Introduction	1
2 Entropy, Symbolization, and Mutual Information	5
2.1 Differential Shannon Entropy	6
2.2 Partition Entropy and Symbolic Dynamics	7
2.3 Symbolization: Discretization of a Continuous State Space	9
2.4 Marginal Spaces and Mutual Information	12
3 Generalizing Geometric Partition Entropy	16
3.1 Preliminaries	17
3.1.1 Ensuring a Bounded Region	17
3.1.2 Handling Repeated Observations within Tolerance	17
3.2 The Voronoi-Based Geometric Partition Entropy	18
3.3 Approximating Geometric Partition Entropy	23

3.3.1	Factor Splitting Algorithms	24
3.4	A Data-Driven SVD Estimator	25
3.5	The Box Estimator	27
3.6	Alternate Measures for the π Estimator	30
3.7	Mutual Information Estimation Using GPE	31
3.7.1	The Partition Intersection Estimator of MI	32
3.7.2	The Box and Voronoi Estimators of MI	33
3.8	Dealing with Negative Estimates of Mutual Information	35
4	Granular Ball Estimator	37
4.1	Motivation	37
4.2	Introduction	38
4.3	Granular-Ball Computing	38
4.4	Generation of Granular Balls	39
4.5	The Adaptive Conditions of Granular-ball Splitting	40
4.5.1	Weighed purity sum of child balls	40
4.5.2	De-overlap between heterogeneous granular-balls	44
4.5.3	An adaptive purity lower bound	44
4.6	Mathematical Estimation of Mutual Information through Granular-	
	Ball Estimator	45
5	Results	52
5.1	High-Dimensional Random Variables	52
5.2	Outputs	54
6	Conclusion	56
A	MATLAB Implementation for High-Dimensional Mutual Infor-	
	mation Estimation	62

List of Figures

2.1	A histogram estimate for a probability distribution from a simple data set $(X, Y) \in \mathbb{R}^2$; the joint entropy $H(X, Y)$ is computed in the joint space, while marginal entropies $H(X)$ and $H(Y)$ are computed from the projected histograms. Figure reproduced from [30].	13
3.1	(a) Given a sample of $N = 500$ unique values (within a tolerance of $\epsilon = 10^{-5}$) lying within a bounded region $\mathcal{D}$, (b) the Voronoi diagram of the data set intersected with the boundary of the region $\mathcal{D}$ provides the optimal partition of $\mathcal{D} \subset \mathbb{R}^d$ for the choice of $K = N$ regions. Coarse-grained estimates of GPE can be obtained for $K = 32$ macrostates by fusing nearly equally represented groups obtained by (c) RPFS or (d) CKM into a partition of supercells. .	22

3.2	(a) Given a sample of $N = 500$ unique values (within a tolerance of $\epsilon = 10^{-5}$) lying within a bounded region $\mathcal{D}$, (b) The Voronoi diagram of the data set intersected with the boundary of the region $\mathcal{D}$ provides the optimal partition of $\mathcal{D} \subset \mathbb{R}^d$ for the choice of $K = N$ regions. Coarse-grained estimates of GPE can be obtained for $K = 32$ macrostates by fusing nearly equally represented groups obtained by (c) RPFS or (d) CKM into a partition of supercells. Figure reproduced from [30].	27
3.3	The RPFS-defined partition using parameter $K = 32$ of the sample from Fig. 3.1(a), where the boxes on the boundary are intersected with $\mathcal{D}$, along with the FS-defined partition for comparison in (c) and (d), respectively. (Source: [30])	29
4.1	Granular-ball iterative splitting and convergence process using k -means clustering.	39
5.1	Comparison of PI, GB, and Exact Gaussian Ground Truth	54

CHAPTER 1

Introduction

Entropy-based analysis of continuous data, most commonly expressed through Differential Shannon Entropy (DSE), has been a cornerstone of information theory since its introduction by Claude Shannon in the final part of his seminal work [1]. However, DSE often provides a poor quantification of uncertainty [2] from data for unknown probability distributions when i) the sample is sparse and/or ii) significant outliers are present. Despite these shortcomings, if the sample size is orders of magnitude larger than the number of histogram bins used to approximate the probability density function (PDF), this approach can still provide useful estimates. However, in higher-dimensional settings, including the estimation of mutual information and many machine learning applications, the sample sizes required rapidly outpace most reasonable computational resources due to the “curse of dimensionality” associated with a changing concentration of measure [4, 5]. The successful application of these concepts was therefore restricted in many ways for several decades, especially in high-dimensional set-

tings, by either the need for a priori knowledge of the underlying distribution for parametric estimation or the requirement for large amounts of data to enable representative histogram approximations of a continuous PDF.

Significant advancements have been made in recent decades with the introduction of the so-called geometric entropy estimators, which are grounded in k -nearest neighbour (kNN) frequency analysis and epitomized by the renowned Kraskov–Stögbauer–Grassberger (KSG) estimator for mutual information [6]. This statistical approach provided a more reliable nonparametric estimator for higher-dimensional measures of conditional mutual information and transfer entropy, particularly for the analysis of data from coupled dynamical systems.

The promise of these advances has been tempered, however, by their considerable computational complexity and the limitations of statistical analysis, which have restricted their successful application to only those systems with a relatively small number of nodes and interactions [7].

Finally, a thorough examination of the popular KSG method indicates that its purported "robustness to outliers" may be problematic in many applications.

Getting large datasets is often impossible, which makes the KSG estimator unreliable. It struggles with small samples and fails to account for outliers, which are often the most important parts of complex systems. Because of these flaws, alternative methods like GPE are necessary, as they are more resilient and accurate when dealing with limited or messy data.

The authors introduced a new approach to entropy estimation from samples in a one-dimensional continuous state space in [22], and named it Geometric Partition Entropy (GPE). That approach was specifically designed to overcome the known limitations of traditional entropy estimators with respect to the handling of outliers, limited sample sizes, and/or atypical distributions. A set of coarse-grained, evenly represented, but unequal-sized macrostates that partition the state space was obtained by using the sample's quantiles. The original GPE

estimate was then defined as what can be described as the partition (or metric) entropy of this quantile-defined partition including known boundaries using the Lebesgue measure.

Although the GPE estimates only agree in the limiting value (as the parameter $K \rightarrow N$) with those of the associated histogram estimates of DSE for samples taken from a uniform distribution, it is generally accepted that entropy is not well-defined in the context of a continuous state space, as values are equivalent up to additive constants and unit conversions [10]. Due to quantisation, a well-defined value can be obtained at the quantum level [11, 12, 13]. However, for macroscopic-scale observations, some coarse-graining is required to provide a meaningful representation of the state space [14]. Under this assumption, it was shown that GPE provides a more consistent and informative entropy measure with respect to variations in parameter choice than histogram-based estimates of DSE, especially when applied to sparse samples from complex distributions [32]. This discrepancy in value will be further explored here, in light of the work of Kolmogorov and Sinai on dynamical systems, where we will find support for the claim that the limiting value of histogram estimates is highly biased and unreliable. At the same time, those of GPE are more informative and based on theoretical limits (though still biased by sample size).

Regardless, GPE’s ability to incorporate the impact of meaningful outliers and to leverage small samples effectively positioned it as a significant advancement in nonparametric entropy estimation.

In addition to being limited to one dimension, the original quantile-based estimator for Geometric Partition Entropy (GPE) struggled to properly account for exactly repeated values. This limitation was shown in [9] to be related to the failure of histogram-based Differential Shannon Entropy (DSE) estimates in adequately handling outliers, and both of these challenges can be further connected through the lens of Kolmogorov theory. The new approach to GPE better

reflects a collection of non-zero local specific volume measures. Thus, geometric partition entropy is now re-imagined in the context of bounded regions in $\mathbb{R}^d$ and presented in a more mature and adaptable way.

Section 2.1 provides a review of several foundational concepts in information theory for continuous state spaces, including the definition of Differential Shannon Entropy (DSE) and the associated concept of partition (or metric) entropy, which leads to the overall process of symbolising data from continuous state spaces while accounting for repeated observations. After that, chapter 3 defines Mutual Information (MI) and introduces a new definition of GPE that incorporates a preprocessing step to handle exactly repeated values properly and leverages the concept of a Voronoi diagram. Several computationally efficient estimators are then defined based on alternative partitioning schemes, which avoid computing the Voronoi diagram. These approximations are then leveraged to define a new class of estimators for the foundational quantity, Mutual Information (MI). In chapter 4, we introduce the concept of Granular Ball Theory and propose a novel Granular Ball -estimator. The proposed estimator utilises granular-ball-based data representation to improve the efficiency and accuracy of entropy estimation, particularly in high-dimensional spaces. Chapter 5 shows the experimental results and performance analysis of the proposed method. The results show that incorporating Granular Ball Theory significantly enhances the performance of the -estimator in higher dimensions, yielding more accurate and robust estimates than traditional estimators.

CHAPTER 2

Entropy, Symbolization, and Mutual Information

The relationship between entropy and mutual information, which is central to modern information theory, is reviewed with a focus on several key concepts with respect to continuous state space estimators that centre around the process of discretizing data on a continuum into a partition of a finite set of macrostates to enable a symbolization of the data. Namely, we introduce the concepts of differential Shannon entropy and the associated partition entropy of Kolmogorov, and then describe symbolization of continuous data, including a process for incorporating repeated values into the symbolization using measure-based partitions. This is followed by an introduction to Mutual Information (MI), including a description of geometric estimators of MI that have recently enabled significant advances in the study of complex systems, data science, and related disciplines. A thorough review of related concepts is necessary to properly place geometric partition entropy within the wider context of entropy-based analysis tools and to provide motivation for its adoption.

2.1 Differential Shannon Entropy

Shannon’s approach to entropy quantification for a continuous pdf, ρ , defined over a bounded finite measure state space, Ω , took the limit as the number of discrete states goes to infinity, transforming the discrete sum from Shannon theory of communication into the integral form

$$H(\rho) = - \int_{\Omega} \rho \log_2(\rho). \quad (2.1)$$

While mathematically reasonable, we will see that fundamental issues can arise when applied to estimation from finite samples. The integral form has since been termed Differential Shannon Entropy (DSE), and historical approaches to its estimation are primarily divided into two categories: parametric and non-parametric estimation.

Parametric estimators are based on the assumption of a priori knowledge of the underlying distribution. Such estimators can often use the integral form directly. They can also provide reliable entropy estimates for known distributions, including a realistic account of outliers. However, their applicability and effectiveness diminish when applied to samples drawn from distributions of unknown functional form. Although many datasets may exhibit characteristics that suggest a known functional form, making such assumptions can lead to highly biased results. Further, the goal of entropy estimation is to gain insight about an underlying distribution, which, in a sense, can render such estimators paradoxical. Thus, while parametric estimators of DSE are valuable in contexts where knowledge of the data is available or assumptions are well warranted, their utility and accuracy are limited in broader applications.

The generalisation of GPE presented here also enables a similar approach to generalise that class of estimators to higher dimensions by using an alternative measure. We review common approaches to the two main non-parametric estimators: 1) histogram-based estimators of DSE and 2) k-nearest neighbour

statistical approaches.

Histogram estimators are essentially a Riemann sum approximation of the integral form of Shannon’s theory, and as their name suggests, they seek to approximate a continuous probability distribution over a bounded region $\mathcal{D} \subset \mathbb{R}^d$ by dividing the region $\mathcal{D}$ into N equal-sized bins (i.e. uniform Lebesgue measure macrostates); the normalized frequencies of observation over these macrostates defines a discrete approximation of the underlying probability distribution (ρ), which we denote by the vector $\mathbf{p}$. This is used to obtain an N -bin estimate of $H(\rho)$ in the discrete symbolic entropy formula

$$H(\mathbf{p}) = - \sum_{i=1}^N p_i \log_2(p_i). \quad (2.2)$$

However, as the dimensionality of data increases, the required number of observations for a representative sample of the state space grows exponentially. This phenomenon, known as the “curse of dimensionality,” presents a significant challenge to traditional histogram-based approaches in high-dimensional information theory applications. Thus, for many modern practical applications of information theory, such estimates are infeasible.

2.2 Partition Entropy and Symbolic Dynamics

Soon after the introduction of Shannon’s theory of entropy, the related concept of partition entropy (also known as metric entropy) was first explored by Kolmogorov as a metric invariant in dynamical systems [28]. Through identification with a Bernoulli process, the adaptation of Shannon’s theory to a distribution of measure over a partition of a measure space was described. This metric-based approach to entropy estimation was shown by Y. Sinai to generalize to a Bernoulli scheme with N possible outcomes and probability vector p in [15], for which the partition entropy is then given by Equation (2.2) for this discrete distribution, connecting the histogram estimate of the entropy of a probability density function

with the entropy of a measurable partition over a probability space. The specific case considered by Sinai [15] consisting of a finite partition $\mathcal{A} = \{A_1, A_2, \dots, A_K\}$ of a measure space, is directly relevant to the formulation of geometric partition entropy. So the author defined the partition entropy limited to this case.

Definition 2.1. *For a bounded set $\mathcal{D} \subset \mathbb{R}^d$, let $(\mathcal{D}, \sigma(\mathcal{A}), \mu)$ be the probability space where $\sigma(\mathcal{A})$ is the σ -algebra generated by a given partition of $\mathcal{D}$ and μ is taken to be any measure for which $\mu(\mathcal{D}) = 1$. If the set $\mathcal{A} = \{A_1, A_2, \dots, A_K\}$ satisfies:*

1. $\bigcup_{i=1}^K A_i = \mathcal{D}$,
2. $\mu(A_i) > 0$ for all $i \in \{1, 2, \dots, K\}$, and
3. $\mu(A_i \cap A_j) = 0$ for all $i \neq j$,

then the partition entropy of $\mathcal{A}$ is defined to be

$$H(\mathcal{A}) = - \sum_{i=1}^K \mu(A_i) \log_2(\mu(A_i)). \quad (2.3)$$

With this definition in mind, the basic quantile partition estimate of GPE for a one dimensional data sample, as described in [22], is simply the partition entropy (using the Lebesgue measure) of a particular partition of a bounded interval from which the sample was drawn. This partition was defined by taking the i/K -th quantiles for $i = 1, 2, \dots, K-1$ together with the bounds of the state space to form a partition of the interval into K unequal size macrostates. To provide additional context for our generalization to higher dimensions, a more in depth look at the theory behind partition entropy as it pertains to countable partitions will be informative.

Sinai [15] that for the case of a general base measure space $(X, \mathcal{B}, \nu)$, where X is a set, $\mathcal{B}$ is a σ -algebra on X , and ν is a measure, one may consider the relative entropy with respect to some region $\tilde{X} \subset X$ and a countable partition $\mathcal{B} = \{B_1, B_2, \dots\} \subset \mathcal{B}$ of $\tilde{X}$ as long as $\nu(\tilde{X}) = 1$. The entropy relative to this

subset is then defined by $H(\mathcal{B})$ using the convergent infinite sum of the form of Eq. (2.3), meaning that the treatment of more complex regions $\mathcal{D}$ consisting of unions of multiple disjoint subregions is also valid for GPE estimation as long as a normalized measure can be defined on each component. Each observation can be considered to be an observed microstate, and a set of K macrostates is sought such that the known microstates are as equally represented as possible.

This connection between Kolmogorov-Sinai entropy and geometric partition entropy remained unexplored at the time of the publication of [22] and [9], however, it provides significant precedent to the concept of GPE more broadly; and when taken together, these two concepts in the symbolization of a continuous state space provide similar insight into the mechanisms of state transitions in complex systems and help bridge the gap between continuous phenomena and their discrete analysis.

2.3 Symbolization: Discretization of a Continuous State Space

Shannon’s introduction of entropy in his “Mathematical Theory of Communication” revolutionised information theory, particularly in its application to digital communication using a discrete set of symbols. This is especially true when data are sparse or when the underlying distribution contains low-probability regions, making it challenging to discretise the continuous state space in a manner that accurately reflects the inherent information content of the data. In such cases, the basic histogram estimator, which relies on the idea of a Riemann sum and partitions the domain into equal-measure bins weighted by frequency, leads to several limitations.

One significant issue is the loss of information about the underlying continuous dynamics due to the coarse-graining procedure. Where the work of Kol-

mogorov and Sinai sought an invariant partition under a given map, the standard histogram approach to entropy estimation naively breaks up the bounded interval into equally spaced subintervals. Depending on the chosen number of bins, K , determining the desired size of the overall symbolic state space, the equal partitioning often oversimplifies the complex variability in such systems.

In order to represent enough detail in this smaller region while using equal-sized partitions, the overall number of states required increases rapidly, resulting in an explosion in the sample size required for any useful frequency-based analysis. This type of simplification through equi-partitioning can significantly affect entropy estimates, and assumes that the only measure of informativity for a macrostate is the frequency of observation. This assumption may be valid under random sampling from a distribution, but when this approach is applied to dynamical systems, this is decidedly not the case, resulting in skewed representations of the system's actual behavior and potentially leading to inaccurate conclusions about information content.

One of the conditions in the definition of partition entropy is that the partition $\mathcal{A} = \{A_1, A_2, \dots, A_K\}$ must be able to generate a σ -algebra on the set $\mathcal{D}$ (i.e. be a nonempty collection of subsets of $\mathcal{D}$ that are closed under compliment, countable unions, and countable intersections). The frequency-based histogram obtained from a data set can be viewed through the lens of measure theory within the probability space $(\mathcal{D}, \sigma(\mathcal{A}), \#X)$, where the set counting measure $\#X(A)$ is defined as giving the number of elements from the discrete set X found within the argument $A \subseteq \mathcal{D}$. Under the usual equi-partitioning of a histogram estimate, if a bin A_i has frequency zero, that implies $\#X(A_i) = 0$. Since $A_i \neq \emptyset$ but has measure zero, it should be removed from the sigma algebra, and a subset $\tilde{\mathcal{D}}$ should be considered after removing this bin. This describes exactly what occurs in the computation of Eq. (2.2), where zero frequency bins are ignored in the sum.

This point is central to understanding the failures of histogram estimates of DSE as the set of utilized symbolic states may not be a σ -algebra on $\mathcal{D}$ if $\bigcup_i A_i \neq \mathcal{D}$, even if $\bigcup_i \#X(A_i) = \#X(\tilde{\mathcal{D}}) = \#X(\mathcal{D}) = 1$. The effective deletion of the empty bins (which is the common practice) essentially ignores part of the dynamics-occupied geometry, assuming the system will never experience a state transition into this geometric space. This is a major pitfall of histogram-based DSE estimators, but it also explains why DSE works well when the data are densely distributed over the region. However, even in such ideal cases, there is another assumption that does not practically hold either, which is that a near-uniform density exists within each state of the partition. The best theoretical scenario for such an assumption to hold is if $(1 \ll k \leq N \rightarrow \infty)$.

Although less problematic, this second assumption is also true for GPE and can be used as an argument for larger values of the parameter K in its estimation. Importantly, though, where larger parameters (K) in DSE converge to the value $\log_2(N)$, this is not true for GPE. In contrast to histogram DSE estimators, the use of quantiles in the original formulation of GPE prevented the unintentional removal of any of the available state space; furthermore could incorporate additional unobserved regions of a known larger state space. This feature, along with the theory supporting KS entropy, lends credence to the claim that the limiting value of the GPE estimator that is observed as the number of macrostates, K , approaches the sample size N , is more informative (though still biased by N) than the limiting values of histogram estimates, which are strongly biased toward $\log_2(N)$.

Similar to the problem that an empty bin poses in histogram estimations of DSE, there is a related challenge in the original discretization process for GPE, which is the proper inclusion of (nearly) repeated values. This will necessarily require the consideration of repeated values within a dataset. While frequency-based histogram estimates struggle to appropriately account for the impact of

outliers due to empty bins, they have no problem accounting for the impact of repeated values. On the other hand, any geometric approach, including both kNN-based approaches and GPE, must address this challenge. The impact on kNN-based methods can be mitigated to a degree for small numbers of repeated values by using a larger k value. Still, this strategy fails when dealing with dynamics that ever display even nearly periodic behaviour.

And, while GPE can deal effectively with near-periodic dynamics, the quantile-based estimator of GPE from [22] fails in the presence of even a small number of exactly repeated values. This concern was, in part, the motivation for the introduction of the Boltzmann-Shannon Interaction Entropy, and so a more complete discussion on this point was presented in [9]. This failure of the original quantile-defined GPE estimator can also be understood in terms of the KS entropy; i.e., in the presence of repeated values, quantiles may coincide, resulting in zero Lebesgue-measure bins. The new approach to estimating GPE in $\mathbb{R}^d$ will now address this important challenge head-on by requiring a tolerance for the inclusion of repeated values (which can default to just above machine precision).

2.4 Marginal Spaces and Mutual Information

The interplay of entropy and Mutual Information (MI) is a foundational relationship of modern information theory. The entropy of a data set serves as a measure of uncertainty, quantifying the spread of a sample and therefore providing a measure of predictability in the phenomena from which it was recorded. Mutual Information builds upon this basic concept by quantifying the extent to which knowledge of one variable reduces the uncertainty (or equivalently increases the predictability) about another [24, 25, 26].

To define MI, we must first introduce several related concepts associated with high-dimensional data sets obtained by concatenating lower-dimensional

state spaces. For example, if $X \subset \mathbb{R}^{d_1}$ and $Y \subset \mathbb{R}^{d_2}$ are equal length sets of observations in d_1 and d_2 dimensions, respectively, then we can consider the data set (X, Y) , consisting of observations in the joint space, $\mathbb{R}^{d_1+d_2}$, where the pairwise associations of observations in X and Y are considered an important feature of the joint data set. One can then estimate the joint entropy of this data in the concatenated space, denoted as $H(X, Y)$, simply by using a $d = d_1 + d_2$ -dimensional histogram with Eq. 2.2; although here is where the sample size required for accurate estimation often prohibits the use of such estimators unless d remains low.

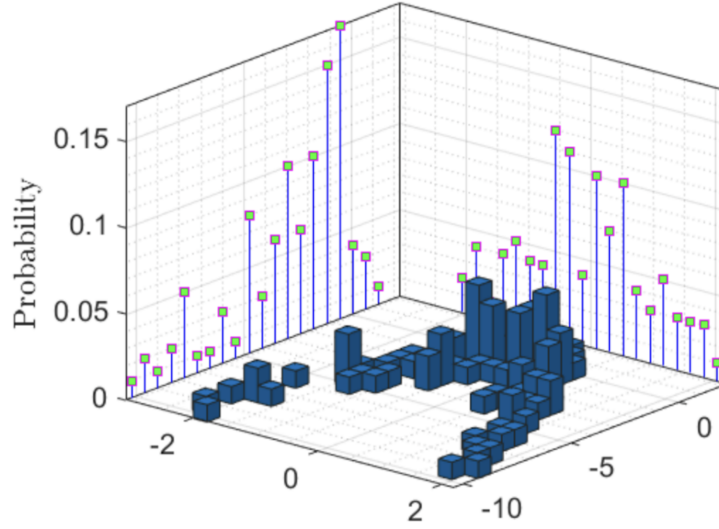


Figure 2.1: A histogram estimate for a probability distribution from a simple data set $(X, Y) \in \mathbb{R}^2$; the joint entropy $H(X, Y)$ is computed in the joint space, while marginal entropies $H(X)$ and $H(Y)$ are computed from the projected histograms. Figure reproduced from [30].

To illustrate, Fig. 2.1 shows a histogram representation of a probability distribution for a low-dimensional data set in the case of $d_1 = d_2 = 1$, where the histogram approximations for the marginal distributions, obtained by projecting the data set (X, Y) onto the marginal spaces X and Y , are shown on the vertical panels. The marginal entropies, $H(X)$ and $H(Y)$, are estimated from these projected histogram approximations of the data set's marginal probability distributions using Eq. 2.2. A histogram estimate for a probability distribution from a simple data set $(X, Y) \in \mathbb{R}^2$; the joint entropy, $H(X, Y)$, would be computed from this histogram in the joint space using Eq. 2.2, whereas the marginal entropies $H(X)$ and $H(Y)$ would be computed using Eq. 2.2 on the one dimensional histograms shown on the vertical planes.

Another important and related concept is the conditional entropy. The entropy of Y conditioned on X is denoted by $H(Y | X)$. It seeks to quantify the remaining uncertainty in the random variable Y (the amount of Information required to encode Y), given that we know the value of another random variable X . In other words, $H(Y | X)$ tells us how uncertain Y is if X is known. If X and Y are independent, knowing X does not provide any information about Y , and $H(Y | X) = H(Y)$. Conversely, if X completely determines Y (though they are not necessarily equal), then $H(Y | X) = 0$ because there is no uncertainty left in Y once X is known.

Using these concepts, the Mutual Information (MI) between two random variables X and Y can be defined in several ways, including:

$$\begin{aligned} I(X; Y) &= H(X) - H(X | Y) \\ &= H(Y) - H(Y | X) \\ &= H(X) + H(Y) - H(X, Y). \end{aligned} \tag{2.4}$$

Based on the theoretical bounds of the component entropies in these definitions, it should be clear that the mutual Information is itself bounded, i.e. $0 \leq I(X; Y) \leq H(X)$, and in particular, it should never be negative.

However, the joint and marginal entropy estimates for data in continuous state spaces are often biased in different ways, e.g., toward the number of bins with non-zero measure, which may differ due to the sample (X, Y) being more sparse in the joint space with respect to its dimensionality as compared with the marginal distributions. This particular manifestation of the curse of dimensionality has been known to yield negative estimates of mutual Information, as the joint entropy is artificially inflated by sparsity and the larger number of zero-frequency bins.

CHAPTER 3

Generalizing Geometric Partition Entropy

A new presentation of geometric partition entropy is now provided that can be applied to data samples drawn from distributions whose domain is a bounded region of $\mathbb{R}^d$. We begin in Sec. [3.1](#) by emphasising the necessity of a well-defined boundary of finite measure for the domain $\mathcal{D} \subset \mathbb{R}^d$ to ensure meaningful probabilities; then, go on to describe the pre-processing step in more detail that ensures the appropriate handling of repeated values. Once these preliminary details are discussed, the Voronoi-based definition of GPE is provided in Sec. [3.2](#). Unfortunately, any actual estimates based on this theoretical definition are computationally expensive, especially in higher dimensions, and the use of irregular partitions in the joint space poses inherent challenges for identifying marginal probability distributions. Thus, we follow the Voronoi-based definition in Sec. [3.2](#) and employ several approximation techniques that address these challenges directly by leveraging alternative partitioning strategies and adapting measures effectively.

Several approximation methods enable the subsequent definition of a class of powerful and efficient mutual information estimators in Sec. [3.3](#), which can incorporate informative outliers in a dynamic and versatile manner.

3.1 Preliminaries

3.1.1 Ensuring a Bounded Region

A bounded (finite measure) region is required for the definition of meaningful probabilities with respect to volumes. In scenarios that lack explicit boundary definitions, approximations of effective limits can significantly influence the estimate of GPE. This is not seen as a failure of GPE, but rather an important feature, specifically, because one goal of this approach is to properly include the impact of outliers, especially when they might be informative.

Ordinarily, the convex hull of a dataset serves as a reliable boundary in many applications, but not necessarily so here. The convex hull uses actual data points to define the boundary, meaning the regions around those particular observations will be underrepresented unless the observations are in fact extremal values within the state space. At the same time, the potentially large regions between these boundary points might become overly emphasized if they are not reasonably expected to be in the true state space. The use of an alpha shape can omit the large unobserved sections of the state space and can be generally advantageous, but these still use data points to define the bounds, likely limiting the proper inclusion of outlier impacts.

3.1.2 Handling Repeated Observations within Tolerance

In addition to a well-defined boundary for a measurable state space, we must also address the known problem of repeated values in the previous approach. In

theory, infinitely accurate observations made on a continuum will result in a zero per cent chance of obtaining exactly repeated values. However, in practice, this is never truly the case. Putting aside more foundational reasons, e.g., machine precision or even the Heisenberg uncertainty principle, there is often a known measurement error that further limits the accuracy of observations.

Thus, any analysis of this type must begin by first defining a tolerance ϵ that determines when observations in the state space should be considered effectively the same, i.e. as a repeated observation of the same microstate. This initial pre-processing step, importantly, does not discretise the state space, but rather redefines the data within the acceptable limits of measurement accuracy to prevent the eventual formation of macrostates below the limits of measurement tolerance.

Thus, for a chosen tolerance ϵ , two points x_i, x_j in a dataset $X \subset \mathcal{D}$ consisting of N points can be considered the same observable microstate (within tolerance) whenever $\|x_i - x_j\|_\infty \leq \epsilon$.

This means that we might end up with $M < N$ unique observed microstates along with an associated vector of the repetition number for each state. Importantly, this prevents any element of a partition of the space from having measure zero, which is required by the formulation of partition entropy put forward by Kolmogorov and Sinai.

3.2 The Voronoi-Based Geometric Partition Entropy

Thus, given a set of observations $X = \{x_n\}_{n=1}^N$ within a bounded (finite measure) region $\mathcal{D} \subset \mathbb{R}^d$, a parameter K , and a tolerance ϵ , we first identify any observations $x_i \in X$ such that there exists at least one $x_j \in X$ with $\|x_i - x_j\|_\infty \leq \epsilon$. A greedy merging process can use centroids of subsets to obtain a set of unique

microstates (within a tolerance of ϵ) along with a corresponding repetition number. After this initial pre-processing step is complete, there are $M \leq N$ unique observed microstates defined on the continuous space $\mathcal{D}$ with an additional frequency of repetition vector of length M .

Similar to the approach in [22], we now seek a partition $\mathcal{A} = \{A_1, A_2, \dots, A_K\}$ of the state space $\mathcal{D}$ such that each subset A_i contains roughly the same number of *unique* observed microstates while ensuring that $\bigcup_i A_i = \mathcal{D}$. This was accomplished in one dimension through the computation of the $\frac{i}{K}$ -th quantiles for $i \in \{1, 2, \dots, K-1\}$, taken together with the bounds of the state space. This step in the process, however, in and of itself, is now a challenging problem with no clear optimal solution due to a lack of a total ordering of the data in higher dimensions.

A path forward is identified by first considering the special case $K = M$ (the number of unique microstates), which has a unique optimal solution provided by the Voronoi diagram of the set of unique observations in $\mathbb{R}^d$ intersected with the boundary of the region $\mathcal{D}$. This results in a geometric partition $\mathcal{V} = \{v_1, v_2, \dots, v_M\}$ of the state space $\mathcal{D}$, where each Voronoi cell $v_m \in \mathcal{V}$ represents the region of the state space that is closest to the unique observation, x_m . One might say that each Voronoi cell, v_m , represents a macrostate of the continuous state space for which the single observed microstate, x_m , is considered the best representative. This particular partition, however, assumes that the data set is, in fact, an optimally chosen independent and identically distributed (i.i.d.) sample from the true underlying distribution. This rather unrealistic assumption can be indicative that using $K = M$ overfits the estimate of the underlying distribution to the particular sample, and so, choosing $K < M$ generally provides a more robust estimate.

In the jargon of measure theory, we consider the probability space $(\mathcal{D}, \sigma(\mathcal{V}), \mu)$ with the σ -algebra generated by the partitioning set of truncated Voronoi cells

$(\mathcal{V})$, which satisfies the assumptions in Def. 1, and the measure μ is taken to be

$$\mu(A) = \frac{N}{m(\mathcal{D})} \frac{m(A)}{\#X(A)}, \quad (3.1)$$

where $m(\cdot)$ is the Lebesgue measure and $\#X(\cdot)$ is the same set counting measure, meaning $\#X(A)$ gives the number of elements of the data set X that fall within the geometric region A . The combined measure μ is representative of the concept of a local measure of *specific volume*, i.e. an inverse density. The way the Voronoi diagram is defined, we need not worry about any sets for which $\#X(A) = 0$, and the multiplicative constant, $\frac{N}{m(\mathcal{D})}$, is chosen to enforce the probability constraint so that $\mu(\mathcal{D}) = \sum_{m=1}^M \mu(v_m) = 1$.

Due to the obvious challenges that can arise with the Lebesgue measure in higher dimensions, e.g., co-linear points, some approximation techniques will employ an alternative to $m(\cdot)$. However, this alternative must still satisfy the definition of a measure. Regardless of the chosen geometric measure, it is important to note that $\#X(\cdot)$ will properly include any repeated values.

In order not to overfit the estimate to the particular data sample, we will want to consider partitions of $\mathcal{D}$ with $K < M$ partition elements. Due to the lack of a total ordering in higher dimensional spaces, it is no longer clear which observations should be grouped with what set of nearby neighbors so that $\mathcal{D}$ is partitioned into K subregions having nearly equal representation among the observed microstates. Importantly, despite this level of uncertainty, it is clear that all such partitions can be represented by members of the $\sigma(\mathcal{V})$, through unions of the various v_m within each cluster. Generally, clustered groups should be formed by geometrically close neighbours in a way that does not result in any of the observations being left out, but this is indeed a problem without a single well-defined solution. However, the benefit of larger values of K means that even greedy clustering of several nearby observations can provide an effective clustering.

In general, a Constrained K-Means (CKM) clustering approach can be used

to great effect however, algorithmic convergence is not guaranteed for smaller K values and this approach can struggle under the specific narrow constraints that each cluster include a minimum of $\lfloor \frac{M}{K} \rfloor$ observations. For this purpose, the author has developed a simple deterministic quantile-based Recursive Prime-Factor Splitting (RPFS) algorithm that provides a reliable and repeatable equitable clustering. The details of this algorithm will be provided in the next subsection where it will enable several approximation techniques.

Regardless of how a set of equitable clusters is obtained, the estimation of GPE proceeds by fusing the corresponding Voronoi cells of the cluster members from each group into larger subregions (i.e. supercells), which, importantly, are members of $\sigma(\mathcal{V})$ with positive measure under μ . Figure 3.2 shows a sample of $N = 500$ points drawn from a uniform distribution along with (b) the Voronoi diagram of this data set whereas panels (c) and (d) show the coarse-grained clustering of this set resulting from the RPFS and CKM algorithms respectively, both using the parameter value of $K = 32$. The corresponding GPE estimates are $H_{RPFS}^{K=32} \approx 4.9356$ and $H_{CKM}^{K=32} \approx 4.9640$.

Since this is a uniform distribution, these estimates are expected to coincide with the histogram estimate of DSE using Eq 2.1, which gives $H_{DSE}^{K=32} \approx 4.9464$. Finally, we note that the supercells should ideally be path connected, but they need not be to provide a useful estimate. Although Voronoi diagrams provide a theoretically accurate definition of GPE, their use presents significant computational challenges. In addition to the cost of constructing the diagram itself, the direct computation of most geometric measures of the resulting irregular shapes can also become problematic in high dimensions using the Lebesgue measure.

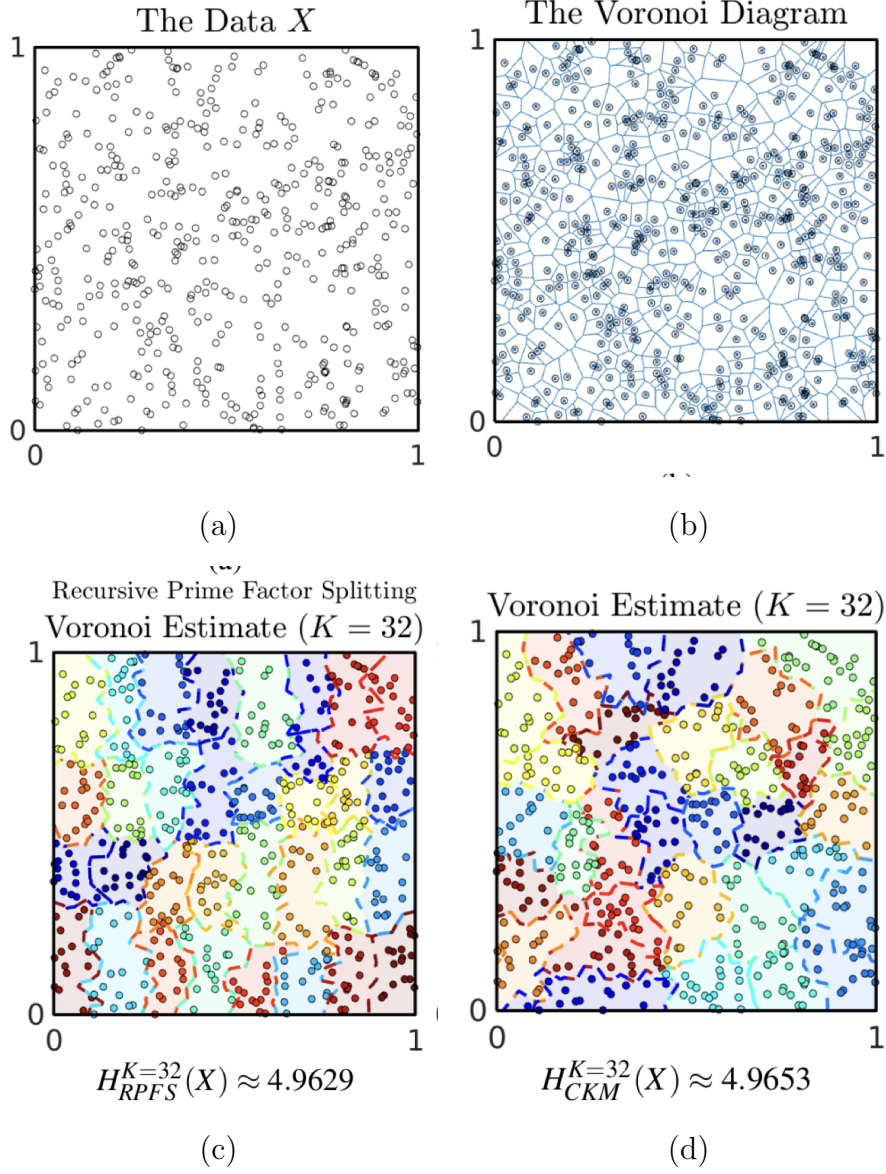


Figure 3.1: (a) Given a sample of $N = 500$ unique values (within a tolerance of $\epsilon = 10^{-5}$) lying within a bounded region $\mathcal{D}$, (b) the Voronoi diagram of the data set intersected with the boundary of the region $\mathcal{D}$ provides the optimal partition of $\mathcal{D} \subset \mathbb{R}^d$ for the choice of $K = N$ regions. Coarse-grained estimates of GPE can be obtained for $K = 32$ macrostates by fusing nearly equally represented groups obtained by (c) RPFS or (d) CKM into a partition of supercells.

To this point, other choices for the probability measure can provide some advantage when used in coarse-grain estimation, but keep in mind the requirements of Kolmogorov, namely that μ must be a measure. For instance, the use of the Chebyshev norm in place of $m(\cdot)$ enables faster computation in high dimensions, but it does not satisfy the definition of a measure in $\mathbb{R}^d$ and leads to inaccuracies.

Furthermore, for the eventual goal of estimating mutual information, the Voronoi diagram partitioning presents a challenge to defining coherent marginal distributions. This underscores the need to avoid relying on Voronoi diagrams and potentially even the Lebesgue measure for fast estimation, particularly in high-dimensional spaces where the performance of both concepts are degraded.

Several approximation techniques have been developed to address these concerns that use different geometric measures and various means of partitioning the state space (with or without a known boundary) to approximate the value given by the Voronoi-based definition successfully, but in both deterministic and more efficient ways.

3.3 Approximating Geometric Partition Entropy

The first several methods of approximation rely on obtaining an initial grouping (or clustering) of the data set into K near-equally represented geometrically distinct subsets. As with the coarse-grained Voronoi estimate, some of the methods that follow can be applied to such clusterings obtained in any way. But, in order to utilize the geometries, we now require an associated partition of the underlying state space, $\mathcal{D}$, to avoid computing the Voronoi diagram.

So, before presenting approximation methods, we first provide a deterministic algorithm and a simplification thereof, each of which not only provides an instance of the desired geometric-based clustering, but also defines a partition of

$\mathcal{D}$ into rectangular boxes defining a set of K macrostates, making several approximations straightforward to compute.

3.3.1 Factor Splitting Algorithms

The Recursive Prime-Factor Splitting (RPFS) algorithm can be applied in either the standard d -dimensional Cartesian basis or subsequently after reorienting the data into a de-meaned Singular Value Decomposition (SVD) vector basis. If using the Cartesian basis, the process begins by first reordering the dimensions of the data set based on the marginal variances in each basis vector direction. The prime factorisation of the parameter K is then computed, which we denote by $K = k_1 \cdot k_2 \cdots k_s$, where the s factors, k_i , are ordered from largest to smallest.

Beginning with the entire set X , at each stage in an iterative process of s steps, the set of current subsets of the data are each split further into k_i subsets along a common basis direction by computing a quantile-based partition of the subset of points projected onto the currently designated basis vector. These partitions use the $\frac{j}{k_i}$ -th quantiles of the subsets of the data for $j = 1, \dots, k_i - 1$ together with the bounds of the current partition of $\mathcal{D}$ restricted to that region of the marginal space. These quantile-based splittings are performed recursively on all current subsets while cycling through the d -dimensional basis vectors in the variance-based order.

In order to avoid zero width quantiles during the various splitting procedures (e.g., dealing with points along a line or plane), the process can be treated with the same tolerance ϵ as before, but applied in the projected spaces, where if the number of unique projected values is smaller than the prime factor, that group can be split into a smaller number of bins but the associated value of K must then be adjusted after the process is completed to accurately reflect the size of the partition.

Perhaps not surprising then, partitioning of this type works best when the

parameter K is taken to be $K = 2^m$ with the integer m being divisible by the dimensionality of the data, d . However, it can be applied for any choice of K , though the existence of large prime factors can lead to inaccuracies, which make estimations based on this form of clustering more sensitive to the choice in parameter K .

The Factor Splitting (FS) algorithm begins by first grouping the set of prime factors of the parameter K into d subsets whose products are either closest to being equal (or that best match the proportion of variances in an SVD basis). We then apply the same quantile-based splitting approach as before, but restricting its application to a single factor in each dimension. This simplifies the bookkeeping of boundaries but adds a minimal increase in the bias toward the first basis direction. Due to the grouping of prime factors into larger dimension-based factors, the FS method will no longer necessarily benefit from the restriction to $K = 2^m$, but similarly the quality of partition will improve with the equality (or alignment with the variances) of the d factors. Despite the potential variation with the parameter K , in our eventual application to Mutual Information estimation, we will choose K in a structured way that largely negates these concerns.

The resulting groups obtained from either the RPFS or FS algorithms can then be used on their own for the Voronoi-based supercell estimate of GPE as described in the previous section, or in the next several approximation schemes.

3.4 A Data-Driven SVD Estimator

In the absence of a known boundary, there is a simple data-driven estimator that can leverage any geometric clustering of the data. Given K clusters, the SVD of each data subset is computed, and the product of the singular values for each cluster is used as a geometric measure of the macrostates. This measure is then divided by the frequency count for each macrostate, as usual, to form an estimate

of local specific volumes, similar to μ in the previous section. This approach generally performs well, but requires the computation of SVD subroutines in addition to some clustering routine prior to its application. In addition, the data-driven bounds may not coincide with the true boundary of the region, leading to large errors. The SVD-defined areas are shown as rectangular regions corresponding to the RPFS-defined clusters for the previous uniform and radially bounded normal samples in Figure 3.2 along with their respective entropy estimates.

It should be noted that the radial boundary shown in Figure 3.2 is more spacious than the data-driven areas estimated by the SVD. So this estimator will consistently overestimate the entropy of such a sample with large unobserved low probability regions of the state space. This sort of discrepancy between data-driven bounds and true estimates will be a recurring theme for any region $\mathcal{D}$ with unobserved portions of the state space included, since GPE estimates are designed to be sensitive to outlier impacts.

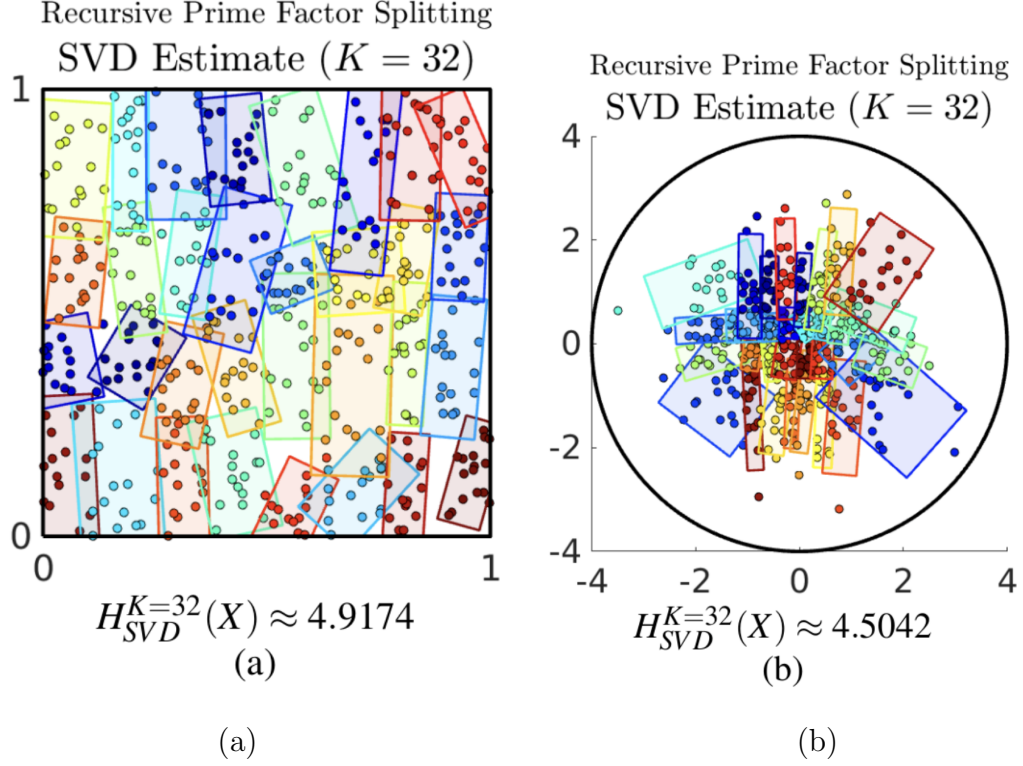


Figure 3.2: (a) Given a sample of $N = 500$ unique values (within a tolerance of $\epsilon = 10^{-5}$) lying within a bounded region $\mathcal{D}$, (b) The Voronoi diagram of the data set intersected with the boundary of the region $\mathcal{D}$ provides the optimal partition of $\mathcal{D} \subset \mathbb{R}^d$ for the choice of $K = N$ regions. Coarse-grained estimates of GPE can be obtained for $K = 32$ macrostates by fusing nearly equally represented groups obtained by (c) RPF or (d) CKM into a partition of supercells. Figure reproduced from [30].

3.5 The Box Estimator

If the boundary for $\mathcal{D}$ is known, these rectangular boxes should be intersected with $\mathcal{D}$ (or the SVD transformation thereof) before computing the geometric measures of the resulting macrostates. Alternatively, if the data is believed to be generally dense in the domain (unlike the radially bounded normal sample),

and $\mathcal{D}$ is mostly convex, the box estimate for GPE will not suffer from using the rectangular box measures directly (i.e., ignoring the boundary). Similarly, if no boundary is known, the resulting boxes can serve as a data-driven proxy for it. However, a similar concern to that referenced in the use of the convex hull arises here in that the boxes will more often include portions of $\mathbb{R}^d$ not in the true bounded state space $\mathcal{D}$ while also using extreme values from the data to define bounds of some boxes.

Whether the boundary is incorporated or not, where the chosen geometric measure of the resulting boxes (potentially intersected with $\mathcal{D}$) is divided by the respective frequency count of the data set within each box, we can estimate the GPE using Equation 2.3. As with the Voronoi definition, since the frequency count will be near uniform by definition (except in the presence of repeated values when projected in the marginal spaces), the measure is largely driven by the geometry. It is thus important to realise that the approximations using the rectangular boxes directly (without intersection with $\mathcal{D}$) may exaggerate the geometry of some macrostates, especially when applied to skewed data. So, if a boundary is known, it should be incorporated. Figure 3.3 show examples of the RPFS-defined partition using $K = 32$ for the same samples of uniform and multivariate normal distributions whereas (c) and (d) do so for the simplified FS-defined partition for comparison.

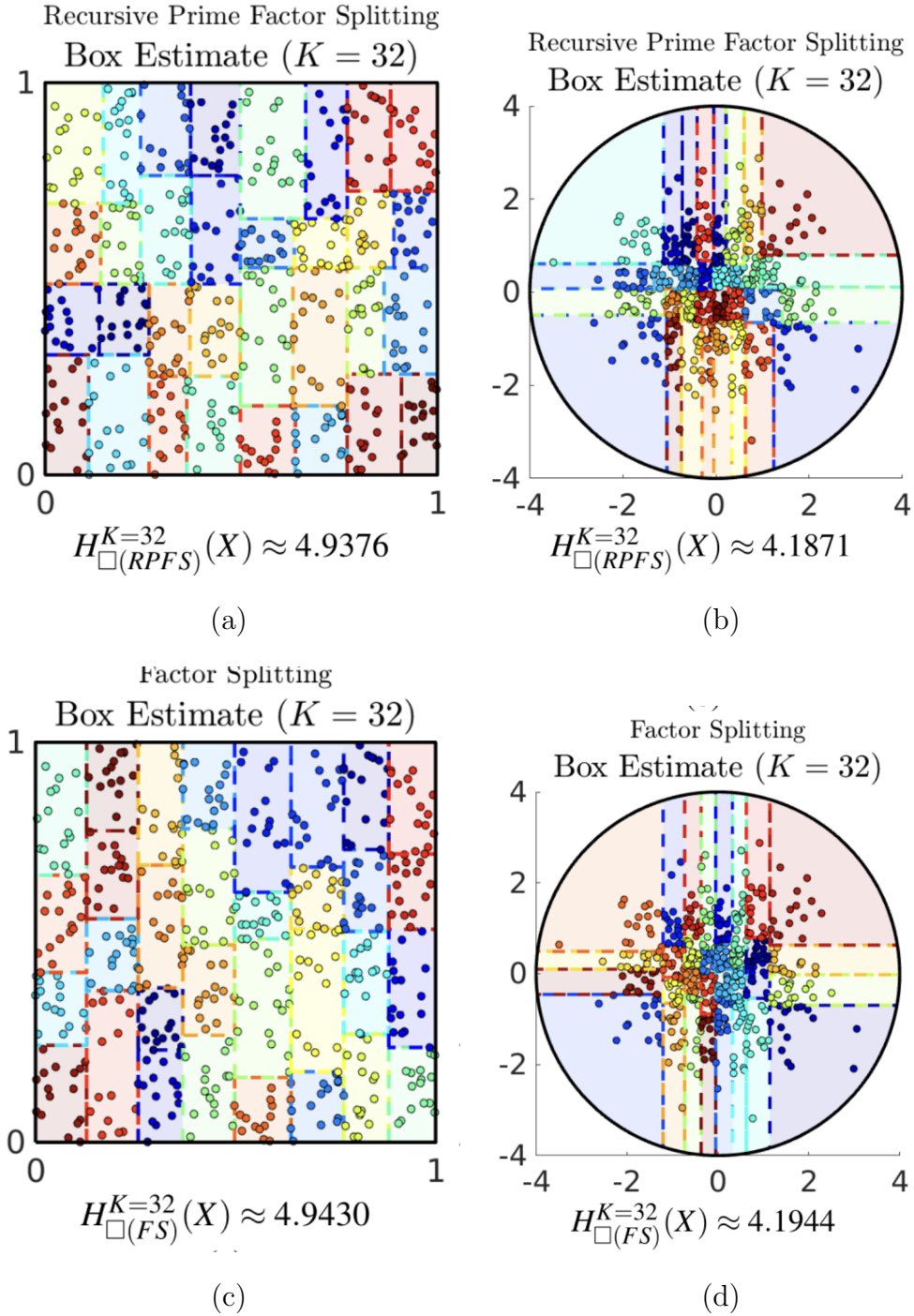


Figure 3.3: The RPFS-defined partition using parameter $K = 32$ of the sample from Fig. 3.1(a), where the boxes on the boundary are intersected with $\mathcal{D}$, along with the FS-defined partition for comparison in (c) and (d), respectively. (Source: [30])

One major hurdle remains in the generalization of GPE for broader applications in information theory. The clusters resulting from the RPFS algorithm, the FS algorithm, or any other geometric clustering approach in a d -dimensional joint space still pose the challenge of extracting marginal distributions. If the marginal spaces of interest are the same dimensionality, we can adapt the above approaches, as will be done in Sec. [3.7.2](#) for the box (and Voronoi) estimator; however, the final approximation method for GPE was developed specifically to enable the development of a more versatile and efficient mutual information estimator, which will also be introduced in the subsequent section. This approach, as applied to GPE estimation for a general parameter value K , tends to exhibit large variation, so it may not be ideal for this purpose. However, its computational complexity and versatility are unmatched by any other estimator and so it is presented in order to provide a clear comparison.

3.6 Alternate Measures for the π Estimator

It is also possible to de-emphasise the impact of outliers by using an alternative measure. If we consider a vector of M geometric measures of the macrostates, denoted by L , and a vector of M macrostate frequencies (including repeated microstate observations), denoted by F , then we can define a measure using any functional form of these two quantities, $g(L, F)$, after appropriate normalisation. Our local specific volume measure $\mu(A) = m(A)\#X(A)/Nm(\mathcal{D})$ is then associated with the functional form $g(L, F) = L/F$. This choice indicates that we place importance on L for quantifying uncertainty, while considering higher values of F in a macrostate to indicate less uncertainty. When applied to equitably represented macrostates, this performs much like the original GPE. However, since this choice is divided by F , it is sensitive to any smaller values in F . While this is not generally a problem in any estimate except the π Estimator, it is also possible

for large numbers of repeated values to affect the accuracy of this choice in other estimators.

Using a measure of the functional form $g(L, F) = L \cdot F$ provides a well-balanced quantification that generally produces results that coincide with histogram-based methods, due to a reduction in the importance of outliers in the π Estimator. Estimates computed using the π partition and this measure may be thought of as a weighted entropy measure specifically designed to reduce the impact of outliers, since low-frequency regions will be much smaller than those with higher frequency. And while other measures address the impact of outliers in different ways, this correspondence further indicates the potentially problematic treatment of outliers by basic histograms.

In addition to these two main choices, we will also consider two other functional forms when the variation of both L and F is large. The functional form $g(L, F) = F \cdot e^{-L}$ allows for partitions with highly repeated values to incorporate the geometry in a meaningful way while bounding its effect on the mostly categorical data represented by highly repeated values; while the use of the functional form $g(L, F) = L \cdot e^{-F}$, as mentioned earlier, this helps include the whole area when using the π Estimator, even if some parts don't have much data or presence.

3.7 Mutual Information Estimation Using GPE

In this section the two computationally efficient approximation schemes for mutual information based on the box and π estimators of GPE are discussed. In particular, the ability to choose various measures within each method makes them versatile. It provides tailored approaches to mutual information estimation for research on complex systems, where outliers may be informative for understanding dynamics.

3.7.1 The Partition Intersection Estimator of MI

The partition intersection (π) approach that was used to partition the state space $\mathcal{D}$ for GPE estimation in began by first defining quantile-based partitions in the marginal space of each dimension. These one dimensional partitions were then intersected within the overall state space, resulting in a rectangular non-uniform grid of boxes that cover the state space $\mathcal{D}$. By design, this enables a straightforward definition of a new mutual information estimator that remains grounded in the basic concept of GPE. However, as before, for the ideas of Kolmogorov and Sinai to apply, we must consider the data set on an alternate bounded region $\tilde{\mathcal{D}}$, consisting of only those bins with nonzero frequency of observation, potentially intersected with the original set $\mathcal{D}$ to remove regions that are known to be outside of the true state space. This means that low probability regions and unobserved portions of the theoretically known state space might be removed, affecting the estimate, similar to what happens with DSE in high dimensions, though much less due to the data-driven bin definitions. Again, this can also be altered by the choice of measure, e.g., $g(L, F) = L \cdot e^{-F}$.

The π estimator for MI, which we denote by $I_\pi(X; Y)$, simply uses the π partitioning approach for estimating the non-conditional entropies in the joint and marginal spaces, then making use of Eq. (4). If X and Y are one dimensional, then the estimates for $H_\pi(X)$ and $H_\pi(Y)$ are simply the quantile-based estimates of GPE from [22] with the newly added weighting of macrostate membership (including repeated values). However, if X and Y reside in higher dimensions, their entropies are computed very much in the same way as the joint entropy, except restricted to the lower dimensional marginal subspaces.

This approach requires the computation of quantile-based partitions in each dimension's marginal space, followed by binning the data within the unequal grid partition that is defined by the intersecting quantiles, where the product of the differences in nearby quantiles provides the geometric measure. In light

of the large number of algorithms for obtaining quantile estimates [?], the π estimation of MI provides an optimal approach with respect to computational complexity, while also being adaptable to any context, e.g., $d_1 \neq d_2$. The added benefit of flexibility in choosing the measure (not restricted to this estimator) also dynamically enables the inclusion of outlier effects across applications.

For example, as before, the use of the alternative multiplicative measure of the form $g(L, F) = L \cdot F$ obtains similar results for MI as the histogram-based DSE and is as consistent as the KSG estimator, supporting the claim that both of these traditional estimators de-emphasise the impact of outliers. Other functions of L and F may also be explored for varying the weighting of outlier impacts.

3.7.2 The Box and Voronoi Estimators of MI

Whenever the data sets X and Y occur in the same underlying state space $\mathcal{D} \subset \mathbb{R}^d$, and $d_1 = d_2 = d$, then the joint space $\mathcal{D} \times \mathcal{D} \subset \mathbb{R}^{2d}$ is the product of two identical copies of the marginal state space. By concatenating the two data sets X and Y in series within the shared space $\mathcal{D}$ (rather than in parallel in the joint space), a single partition of $\mathcal{D} \subset \mathbb{R}^d$ can be computed using either of the factor splitting algorithms (RPFS or FS) on the combined data set. Similarly, the Voronoi diagram of the concatenated data set can be constructed to obtain $2M$ cells (or fewer after further adjustment for repetition across X and Y) within the single copy of the state space. Subsequent geometric clustering, such as CKM, can lead to a shared partition of the marginal space $\mathcal{D}$, however we chose to focus on the box approximation due to its scaling in higher dimensions.

It is important to note that the macrostates defined in this way will now not necessarily have nearly equal numbers of points from both X and Y , but rather an equal number of points from the combined data set. Thus, it is reasonable to expect similar challenges to those found with the π estimator in that large variances in frequency counts from X or Y can lead to poor estimates. For this

reason, which is especially true in higher dimensions, it may be best to use a relatively small parameter value, e.g., $k = 3, 4, 5$, or 6 in each dimension; and, although a bipartition in each dimension ($k = 2$) may struggle to capture detail in lower dimensional applications, it performs surprisingly well as the dimension increases. Further, this approach scales well with comparison across dimensions, since the overall number of bins used will scale exponentially with dimension, i.e. $K = k^d$. They will generally use $k = 6$, except for low-dimensional examples, e.g., $d_1 = d_2 = 1$, where finer-grained detail will be useful.

Since the computation of the Voronoi diagram and even the RPFS-partitioning procedure become more intensive in higher dimensions, we restrict our consideration to the simpler Factor Splitting (FS) approach to define a box estimator for MI. Using the partition obtained from the FS algorithm applied to the combined data $(X; Y) \subset \mathbb{R}^d$ (in series), the joint entropy $H(X, Y)$ is estimated using the symbolised joint states in (X, Y) where the products of the geometries in $\mathbb{R}^d$ for each element of the corresponding symbols in X and Y are divided by the joint symbol frequencies. The marginal entropies $H(X)$ and $H(Y)$ are then computed using the box estimator with the same FS-defined partition of $\mathcal{D}$. Using these component entropy estimates, the box estimator of mutual information follows from Eq. [2.3](#).

One final point to make on this approach is that if the mutual information is high and/or a small enough parameter k is chosen, the FS partitioning procedure will be more likely to include large portions of low probability or even unobserved state space when compared with the π partitioning approach (unless using the $g(L, F) = L \cdot e^{-F}$ form of measure).

3.8 Dealing with Negative Estimates of Mutual Information

By definition, the Mutual Information should always be non-negative. Many MI estimators, but especially the KSG and other knn-based approaches, still give negative values often due to the variation in biases of the marginal and joint entropy estimates. When all observations are considered to be of equivalent informational value, frequency-based measures are biased in predictable ways, and so in general, these negative outcomes are often small in size, and it has been common practice to simply set negative values equal to zero when they arise. However, when leveraging the geometry in a way that exaggerates the role of outliers, those differences in bias between marginal and joint entropy estimates can become much larger. Thus, we find that both the I_π and $I_\square$ estimators can take on much larger magnitude negative values of MI, which is problematic.

We therefore must understand when these challenges arise and provide an adaptive solution. There are two cases to consider: the first is where the outlier impact becomes exaggerated in the joint space due to the curse of dimensionality; and the second arises in the presence of highly repeated values, where the data may be better considered as categorical rather than truly on a continuum. In the second, detailed analysis of the change in entropy with increased tolerance ϵ , this can be informative here. But, regardless of its cause, whenever negative estimates are encountered in either the π or box estimators, instead of artificially setting those values to zero, we dynamically alter the measure μ to be of the form $g(L, F) = F \cdot e^{-L}$.

Recall that the frequency vector F is for a set of (unequal measure) geometrically defined bins, meaning the informational value of each observation is proportional to the size of the bin in which it is found, with larger bins implying more informative observations. The multiplication by e^{-L} in this measure then

slightly reduces the importance of the geometry. This reduction in the importance of geometry, coupled with increased attention to frequency (though still defined on non-equal-measure partitions), results in more meaningful estimates for highly repeated values while also reducing the impact of large geometries in the joint space.

The MI recomputed using a measure of the form $g(L, F) = F \cdot e^{-L}$ results in a smooth transition from estimates using $g(L, F) = \frac{L}{F}$, while remaining positive in such cases. To deal with the negative estimates of Mutual Information, we proposed the Granular Ball estimator, which we discuss in detail in chapter [4](#).

CHAPTER 4

Granular Ball Estimator

4.1 Motivation

Although π -partition estimators improve on traditional histograms by using quantiles to achieve more balanced data distribution, they remain constrained by their rigid rectangular boundaries. These fixed partitions often fail to capture complex or irregular data patterns and become increasingly inefficient in higher dimensions, leading to unstable performance.

To overcome these limitations, we incorporate Granular-Ball Computing (GBC), which replaces the inflexible grid structure with adaptive geometric “granular balls.” This approach handles empty space and outliers more robustly, and delivers more accurate and stable mutual information estimation for datasets with non-uniform or irregular distributions.

4.2 Introduction

Granular computing is a scalable, efficient and robust method that is very similar to the way the human brain thinks. Therefore, granular computing represents a very viable research area, as indicated by the following quote defined by Yao [18]: “As an emerging field of study, the granular Computing (GC) is both new and old”. The term “granular computing” was first proposed in 1997, but the ideas and principles of granular computing have been widely studied for years within various fields.

Most methods which apply grouping, classification and clustering to analyze and solve problems are within the scope of granular computing. Namely, granular computing uses simple, low-cost, satisfactory approximate solutions rather than exact solutions. Besides, it uses the compatibility of inaccurate, incomplete, uncertain, and massive information to achieve tractable, robust, low-cost intelligent systems and intelligent controllers that can describe the real world better [19].

4.3 Granular-Ball Computing

In the literature [20, 21], Granular-Ball Computing (GBC) is introduced into existing classification models to generate granular balls with different radii covering the dataset.

Definition 4.1. *Let X be a nonempty set of objects, $X = \{x_1, x_2, \dots, x_M\}$, where M is the number of sample points in X . The center point P and radius r of X are defined as follows $P = \frac{1}{M} \sum_{i=1}^M x_i$, $r = \frac{1}{M} \sum_{i=1}^M d(x_i, P)$, where $d(x_i, P)$ denotes the Euclidean distance between x_i and P . Thus, P represents the center of gravity of all sample points in X , and r is the average distance of all samples from the center point P .*

4.4 Generation of Granular Balls

In granular-ball computing, the granular ball GB is generated based on the k -means clustering algorithm. The generated granular ball is defined as the neighborhood with center point P and radius r , and it is labeled according to the most frequently occurring class label in X . The purity of a granular ball is defined as the percentage of samples having the same label as that of the granular ball. Granular balls are generated iteratively according to a predefined purity threshold.

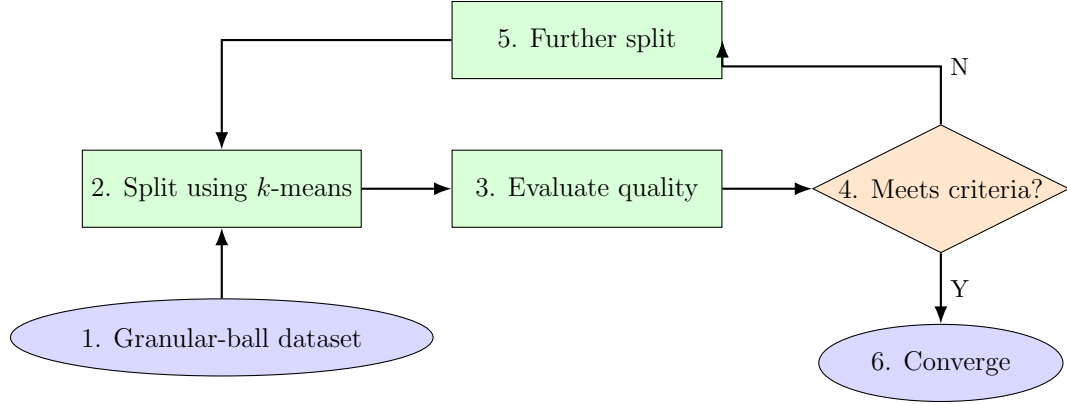


Figure 4.1: Granular-ball iterative splitting and convergence process using k -means clustering.

The generation process of granular balls is illustrated above. Initially, the entire dataset is considered as a single granular ball. Since the purity of this ball is generally low, it usually cannot satisfy the stopping criterion. Therefore, the k -means clustering algorithm is employed to divide the current granular ball into k sub-balls. The purity of each generated sub-ball is then calculated, and it is checked whether the quality requirement is satisfied.

If the purity of every sub-ball reaches the predefined threshold value, the splitting process terminates. Otherwise, those sub-balls whose purities remain below the threshold are further divided iteratively until all generated granular

balls satisfy the stopping condition. Finally, the granular-ball computing process converges and the resulting granular balls are obtained.

‘From the overall process of granular-ball computing, it can be observed that, for a given purity threshold, the radius of each granular ball can be adaptively determined, while the number of generated balls is also automatically obtained. As the splitting process continues, the average purity of the granular balls increases and the decision boundary becomes progressively clearer. Consequently, regardless of the complexity of the data distribution, sufficiently fine granulations (granular balls) can effectively characterize the decision boundary of the dataset.

4.5 The Adaptive Conditions of Granular-ball Splitting

To realise the adaptive generation of granular balls, three conditions are as follows: whether the weighted purity sum of child balls for each granular ball increases or not, whether there is overlap between any pair of heterogeneous granular balls, and whether each granular ball reaches to the lower bound of purity, i.e., the purity of the initial granular ball of the whole data set.

4.5.1 Weighed purity sum of child balls

The purity is designed to measure the quality of a granular ball. A direct idea is to design an indicator to measure the purity of the child balls. Then, whether a granular ball should be split or not is determined by whether the purity of its child granular balls becomes larger than itself. Considering that the more samples in the ball, the more important the ball is, the weighed purity sum of the child balls is used to measure the child balls’ purity, as shown in Definition 2.

Definition 4.2. *Given granular-balls A and its child balls*

$$A_i (i = 1, 2, \dots, k),$$

where

$$\bigcup_{i=1}^k A_i = A$$

and

$$\bigcap_{i=1}^k A_i = \emptyset.$$

$|\cdot|$ denotes the number of elements in a set. k denotes the number of classes in A . A_i^l denotes a set consisting of those samples whose label is equal to l , and A_i^* , i.e., $l = *$, represents the set consisting of those samples in the majority class in the A_i .

The weighted purity sum W of the child balls of A can be defined as

$$W = \frac{|A_1|}{|A|} \times \frac{|A_1^*|}{|A_1|} + \frac{|A_2|}{|A|} \times \frac{|A_2^*|}{|A_2|} + \dots + \frac{|A_k|}{|A|} \times \frac{|A_k^*|}{|A_k|}. \quad (4.1)$$

At the same time, the majority samples in A are equal to the sum of the majority samples in all the child balls. From the Eq. [4.1](#), we get

$$\sum_{i=1}^k A_i^* = \sum_{i=1}^k A_i^l = A^*. \quad (4.2)$$

From the Eq. [4.1](#), we can easily get

$$\begin{aligned} W &= \frac{\sum_{i=1}^k |A_i^*|}{|A|} \\ &= \frac{|A^*|}{|A|} \\ &= T. \end{aligned}$$

So,

$$W = T. \quad (4.3)$$

[2] When $\exists A_i \subset A$ and $\text{label}(A) \neq \text{label}(A_i)$, the majority sample in A are not necessarily the majority sample in all child balls. Assuming $A_m (m =$

$1, 2, \dots, k$) has a different label with A , we can get

$$|A_m^*| > A_m^l. \quad (4.4)$$

In addition, the samples labeled l in the parent ball equal to the sum of those in all child balls, we have

$$A^l = \sum_{i=1}^k A_i^l. \quad (4.5)$$

Combining with Eq. 4.4 and Eq. 4.5, we get

$$\sum_{i=1}^k |A_i^*| = |A_1^*| + |A_2^*| + \dots + |A_m^*| + \dots + |A_k^*| \quad (4.6)$$

$$> A_1^l + A_2^l + \dots + A_m^l + \dots + A_k^l \quad (4.7)$$

$$= \sum_{i=1}^k A_i^l \quad (4.8)$$

$$= A^l. \quad (4.9)$$

From the 4.9 and the 4.2, we can easily get

$$W = \frac{\sum_{i=1}^k |A_i^*|}{|A|} \quad (4.10)$$

$$> \frac{|A^*|}{|A|} \quad (4.11)$$

$$= T. \quad (4.12)$$

$$W > T. \quad (4.13)$$

Based on the definition 4.2, Theorem 4.1 is proposed to describe the condition of that a granular-ball should be split.

When W is greater than T , it means the label of some child balls is different from the parent ball, i.e., the minority samples in the parent ball become the majority samples in some child balls. It can be concluded that the weighted purity sum of the child balls will be greater than the purity of the parent ball,

and the number of correctly classified samples will increase. When W equals T , it represents a special case that the parent ball and all child balls have the same label.

Theorem 4.1. *Given a granular-ball A , whose label is denoted by $\text{label}(A)$ and purity by*

$$T = \frac{|A^*|}{|A|},$$

and its child granular-ball

$$A_i (i = 1, 2, \dots, k),$$

where k is the number of the child granular-balls. $\text{label}(A) = L$. W represents the weighted purity sum of A_i .

(1) $\forall A_i \subset A$, if $\text{label}(A) = \text{label}(A_i)$, then $W = T$;

(2) $\exists A_i \subset A$, if $\text{label}(A) \neq \text{label}(A_i)$, then $W > T$.

Proof:

[1] When $\forall A_i \subset A$ and $\text{label}(A) = \text{label}(A_i)$, the majority samples in A are also the majority samples in all child balls, so we have

$$A^* = A^l. \tag{4.14}$$

So,

$$\begin{aligned} W &= \frac{\sum_{i=1}^k |A_i^*|}{|A|} \\ &= \frac{|A^*|}{|A|} \\ &= T. \end{aligned}$$

[2] When $\exists A_i \subset A$ and $\text{label}(A) \neq \text{label}(A_i)$, we get

$$W > T. \tag{4.15}$$

4.5.2 De-overlap between heterogeneous granular-balls

For the problem of granular-balls overlap, it is necessary to further detect whether there is heterogeneous granular-balls and further split and refine the overlapping granular-balls to make the decision boundary clearer. In order to improve efficiency, the next round of overlap detection only needs to traverse the child granular-balls of the granular-balls that have overlapped.

The boundary overlap problem of heterogeneous granular-balls can be defined as the following model:

$$\begin{aligned} \exists \|c_i - c_j\| \leq \|r_i + r_j\|, \quad \text{for } i, j \in \{1, 2, \dots, m\} \\ \text{s.t. } \text{label}(GB_i) \neq \text{label}(GB_j). \end{aligned}$$

Among them, c_i represents the center of the i th granular-ball, r_i represents the radius of the i th granular-ball, and m is the total number of granular-balls. In addition, the effect of the condition “ $\text{label}(GB_i) \neq \text{label}(GB_j)$ ” is to concentrate the boundary overlap problem between heterogeneous granular-balls, and reducing the cost of computing and analysis of the problem.

The overlap between the same kind of granular-balls will not affect the decision boundary.

4.5.3 An adaptive purity lower bound

In addition, the purity of the granular balls should have an adaptive lower bound. The lower bound is the proportion of the initial majority of samples of the total sample, that is, the purity of the initial granular-ball.

For minority samples, the proportion of incorrectly classified samples can be considered as the noise rate. That is, the purity of all granular balls must be greater than the purity of the initial granular ball.

The optimisation goal of the adaptive granular-ball generation method can be expressed as

$$\begin{aligned}
& \min \quad \lambda_1 \cdot \frac{n}{\sum_{j=1}^m X(GB_j)} + \lambda_2 \cdot m, \\
& \text{subject to} \\
& \quad \text{quality}(GB_j) > T_0,
\end{aligned} \tag{4.16}$$

$$W(GB'_j) > \text{quality}(GB_j),$$

$$\|c_i - c_j\| > \|r_i + r_j\|,$$

$$(i, j \in [1, m], \text{label}(GB_i) \neq \text{label}(GB_j)).$$

where λ_1 and λ_2 are the corresponding weight coefficients and $k < n$, and c_i, r_i represents the center and radius of GB_i respectively. T_0 denote the adaptive purity lower bound of granular-balls, and GB'_j represents the child granular-balls of GB_j . In addition, $\text{label}(GB_i)$ and W are mentioned above.

4.6 Mathematical Estimation of Mutual Information through Granular-Ball Estimator

This work presents an estimation framework for granular-ball-based mutual information in high-dimensional data. The proposed approach replaces traditional fixed geometric partitions with adaptive granular balls. Each granular ball is characterised by its centre, radius, purity, and local sample density, enabling a multi-granularity representation of complex distributions. Adaptive splitting criteria govern the generation of granular balls based on weighted purity improvement, overlap reduction between heterogeneous granular balls, and an adaptive purity lower bound. These mechanisms ensure that the partitioning process re-

mains both informative and computationally efficient while preserving meaningful decision boundaries.

We constructed granular-ball partitions and then estimated local probability masses directly from the sample distribution. Marginal and joint probability estimates are then obtained through granular balls in the respective feature spaces, leading to a granular-ball mutual information estimator formulated as

Let the dataset be

$$D = \{(x_i, y_i)\}_{i=1}^N,$$

where

$$x_i \in \mathbb{R}^{d_x}, \quad y_i \in \mathbb{R}^{d_y}.$$

The objective is to estimate the mutual information between the random variables X and Y using adaptive granular-ball partitions.

Step 1: Granular-Ball Representation

A granular ball is defined as

$$GB = \{x_i \mid i = 1, 2, \dots, N\}.$$

For a granular-ball GB_k ,

$$GB_k = (c_k, r_k),$$

where:

- c_k is the centre,
- r_k is the radius.

Step 2: Centre of a Granular Ball

For a granular-ball containing n_k samples,

$$GB_k = \{x_1, x_2, \dots, x_{n_k}\},$$

its center is

$$c_k = \frac{1}{n_k} \sum_{i=1}^{n_k} x_i.$$

This gives the mean location of all samples within the granular ball.

Step 3: Radius of a Granular Ball

The radius is defined as the average distance of samples from the centre:

$$r_k = \frac{1}{n_k} \sum_{i=1}^{n_k} \|x_i - c_k\|.$$

The radius measures the local spread or compactness of the granular ball.

Step 4: Adaptive Granular-Ball Splitting

The entire dataset initially forms one large granular ball.

This ball is recursively divided according to adaptive conditions.

Suppose a parent ball A is divided into child balls

$$A_1, A_2, \dots, A_k,$$

such that

$$\bigcup_{i=1}^k A_i = A, \quad \bigcap_{i=1}^k A_i = \emptyset.$$

Step 5: Purity of a Granular-Ball

For a granular-ball A_i ,

$$\text{Purity}(A_i) = \frac{|A_i^*|}{|A_i|},$$

where

- $|A_i|$ = total samples in A_i ,
- $|A_i^*|$ = majority-class samples in A_i .

Purity measures the homogeneity of the granular ball.

Step 6: Weighted Purity Sum

To determine whether splitting is beneficial, define the weighted purity sum:

$$W = \sum_{i=1}^k \frac{|A_i|}{|A|} \cdot \frac{|A_i^*|}{|A_i|}.$$

Simplifying,

$$W = \sum_{i=1}^k \frac{|A_i^*|}{|A|} = \frac{\sum_{i=1}^k |A_i^*|}{|A|}.$$

Step 7: Splitting Criterion

Let

$$T = \text{Purity}(A)$$

be the purity of the parent ball.

Then:

- If $W > T$: splitting improves the representation.
- If $W = T$: the split is redundant.

Thus, splitting continues only when purity increases.

Step 8: Overlap Condition

For two granular balls GB_i and GB_j ,

$$GB_i = (c_i, r_i), \quad GB_j = (c_j, r_j),$$

overlap occurs when

$$\|c_i - c_j\| \leq r_i + r_j.$$

Only heterogeneous granular balls are considered:

$$\text{label}(GB_i) \neq \text{label}(GB_j).$$

This improves boundary discrimination.

Step 9: Adaptive Purity Lower Bound

Define the initial purity threshold.

$$T_0 = \frac{\text{Majority Samples}}{\text{Total Samples}}.$$

Every granular ball must satisfy

$$\text{quality}(GB_j) > T_0.$$

This prevents low-quality granular balls.

Step 10: Optimisation Model

The adaptive granular-ball generation process is formulated as

$$\min \lambda_1 \left(\frac{\sum_{j=1}^m X(GB_j)}{n} \right) + \lambda_2 m,$$

subject to

$$\text{quality}(GB_j) > T_0,$$

$$W(GB'_j) > \text{quality}(GB_j),$$

$$\|c_i - c_j\| > r_i + r_j,$$

$$\text{label}(GB_i) \neq \text{label}(GB_j).$$

where:

- m = number of granular balls,
- $X(GB_j)$ = information contribution of GB_j ,
- λ_1, λ_2 = balancing coefficients.

Step 11: Probability Estimation using Granular-Balls

Suppose:

- total samples = N ,

- granular-ball GB_k contains n_k samples.

Then the local probability estimate is

$$p_k = \frac{n_k}{N}.$$

This estimates the probability mass represented by each granular ball.

Step 12: Marginal Probability Estimation

Construct granular balls separately in the X -space and Y -space.

For the i -th marginal ball of X ,

$$p(x_i) = \frac{n_i^X}{N}.$$

Similarly, for the j -th marginal ball of Y ,

$$p(y_j) = \frac{n_j^Y}{N}.$$

Step 13: Joint Probability Estimation

Construct granular-balls in the joint space (X, Y) .

If the joint granular-ball contains n_{ij} samples, then

$$p(x_i, y_j) = \frac{n_{ij}}{N}.$$

Step 14: Mutual Information Formula

The classical mutual information definition is

$$I(X; Y) = \sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(x_i)p(y_j)}.$$

Substituting the granular-ball probability estimates,

$$p(x_i, y_j) = \frac{n_{ij}}{N}, \quad p(x_i) = \frac{n_i^X}{N}, \quad p(y_j) = \frac{n_j^Y}{N},$$

we obtain

$$I(X; Y) = \sum_{i,j} \frac{n_{ij}}{N} \log \left(\frac{\frac{n_{ij}}{N}}{\left(\frac{n_i^X}{N}\right)\left(\frac{n_j^Y}{N}\right)} \right).$$

Step 15: Simplification

Simplifying the denominator,

$$\left(\frac{n_i^X}{N}\right) \left(\frac{n_j^Y}{N}\right) = \frac{n_i^X n_j^Y}{N^2}.$$

Therefore,

$$I(X; Y) = \sum_{i,j} \frac{n_{ij}}{N} \log \left(\frac{n_{ij} N}{n_i^X n_j^Y} \right).$$

Hence, the granular-ball mutual information estimator becomes

$$I(X; Y) = \sum_{i,j} \frac{n_{ij}}{N} \log \left(\frac{n_{ij} N}{n_i^X n_j^Y} \right),$$

which provides an adaptive geometric estimation of mutual information using granular-ball partitions.

CHAPTER 5

Results

5.1 High-Dimensional Random Variables

Considering how these estimators scale with dimension, we provide a simple experiment in the rare case of a known ground truth. For each choice in dimension, d , we generate two samples X and Y of size N , where the i -th dimensions, x_i and y_i , are functions of the same random variable $t_i \sim \mathcal{N}(0, 1)$. Specifically, we define $x_i(t_i)$ and $y_i(t_i)$ as cubic polynomial functions of t_i as follows:

$$x_i(t_i) = a_0 + a_1 t_i + a_2 t_i^2 + a_3 t_i^3$$

$$y_i(t_i) = b_0 + b_1 t_i + b_2 t_i^2 + b_3 t_i^3$$

where the coefficient vectors $\mathbf{a}, \mathbf{b} \sim \mathcal{N}(0, 1)$ are generated randomly in each iteration. This procedure results in high mutual information that scales linearly with the dimension.

Considering Gaussian i.i.d. random variables, the parametric estimator of

mutual information can be defined as:

$$I(X; Y) = \frac{1}{2} \log \left(\frac{\det(\Sigma_X) \det(\Sigma_Y)}{\det(\Sigma)} \right) \quad (5.1)$$

where Σ represents a covariance matrix for the random variables whose (i, j) -th element is given by:

$$\Sigma_{ij} \equiv \overline{x_i x_j} - \overline{x_i} \overline{x_j}. \quad (5.2)$$

The subscripted covariance matrices in Equ. (6.1) are those defined in the appropriate marginal distributions, i.e. for X and Y , while the unsubscripted matrix represents the full $2d \times 2d$ dimensional covariance matrix.

5.2 Outputs

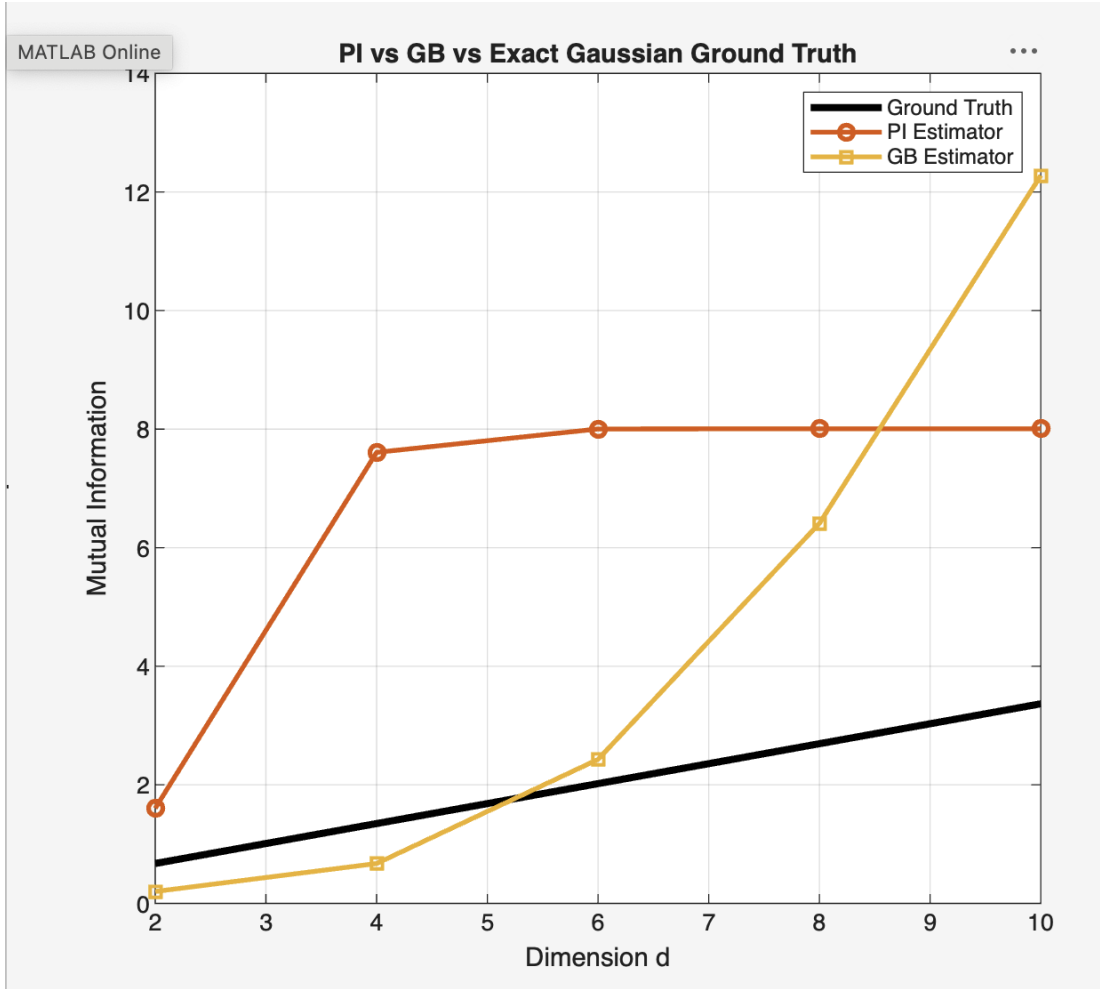


Figure 5.1: Comparison of PI, GB, and Exact Gaussian Ground Truth

The x -axis represents the data dimensionality d , ranging from 2 to 10. At the same time, the y -axis (Mutual Information) denotes the estimated mutual information values obtained using the PI estimator, GB estimator, and the exact Gaussian ground truth.

Exact Gaussian Ground Truth:

The black curve represents the theoretical Gaussian mutual information, which increases gradually with dimensionality. It shows the analytical relation-

ship between dimension and mutual information for correlated Gaussian variables, where the mutual information grows as additional dependent dimensions are introduced.

PI Estimator:

The PI estimator exhibits rapid saturation. After a sharp increase at lower dimensions, the estimated mutual information remains nearly constant around 8 despite further increases in dimensionality. It suggests that the PI estimator loses sensitivity in higher-dimensional spaces and is unable to accurately capture the ongoing increase in mutual information due to sparsity effects and fixed-partition limitations in high-dimensional estimation.

GB Estimator:

Whereas the GB estimator demonstrates a steadily increasing trend as the dimension increases. The GB estimator follows the general growth pattern of the theoretical ground truth more closely, indicating improved sensitivity to high-dimensional dependency structures. However, at larger dimensions, the GB estimator begins to overestimate the mutual information relative to the exact Gaussian value.

Overall, the figure illustrates the challenges of high-dimensional mutual information estimation. The PI estimator appears numerically stable but suffers from early saturation and under-representation of high-dimensional dependency information. The GB estimator, on the other hand, is more adaptive to dimensional growth and better captures the increasing information content of the data.

CHAPTER 6

Conclusion

Mutual information is a key concept in information theory that measures the nonlinear dependence between random variables. Estimating mutual information becomes increasingly difficult as dimensionality increases, mainly due to the curse of dimensionality and the limitations of traditional partition-based methods. The Partition intersection estimator uses quantile-based partitioning of the state space to estimate mutual information, offering interpretability and computational simplicity in low- to moderate-dimensional settings. However, its performance declines considerably as the dimensions increase due to increased partition complexity. To overcome these limitations, this thesis developed a Granular Ball (GB) estimator for mutual information.

The proposed estimator naturally adapts to skewed data distributions, reducing the sparsity issues commonly encountered in high-dimensional settings. An experimental study based on a smooth latent-variable model demonstrates that the granular-ball estimator is more sensitive to dimensional growth and bet-

ter captures dependency structures than the conventional Partition intersection estimator. While the Partition Intersection estimator saturates as dimensionality increases, the granular-ball estimator maintains a steadily increasing trend that more closely follows the theoretical behaviour of the mutual information. The results show that the proposed method provides improved adaptability, robustness, and estimation stability, making it a promising framework for practical mutual information estimation tasks. In conclusion, this thesis contributes to the field of mutual information estimation by identifying the limitations of the Partition Intersection estimator in higher dimensions and introducing a novel Granular Ball estimator that achieves superior performance in such settings. The proposed approach offers both theoretical significance and practical potential for high-dimensional dependence analysis.

Bibliography

- [1] Shannon, C. E., *A Mathematical Theory of Communication*, Bell System Technical Journal, 1948.
- [2] Paninski, L., *Estimation of entropy and mutual information*, Neural Computation, 2003.
- [3] Beaumont, O., Kermarrec, A. M., Marchal, L., and Rivière, E., *VoroNet: A scalable object network based on Voronoi tessellations*, Proceedings of the IEEE International Parallel and Distributed Processing Symposium (IPDPS), IEEE, 2007.
- [4] Sricharan, K. et al., *Ensemble estimators for multivariate entropy estimation*, IEEE Transactions on Information Theory, 2013.
- [5] Goldfeld, Z. et al., *Estimating information-theoretic quantities with neural estimators*, 2018.
- [6] Kraskov, A., Stögbauer, H., and Grassberger, P., *Estimating mutual information*, Physical Review E, 69(6), 066138, 2004.

- [7] Zhao, Y. et al., *Analysis of KSG estimator limitations in high-dimensional systems*, 2019.
- [8] Christopher Tyler Diggans and Abd AlRahman R. AlMomani, “Geometric partition entropy: Coarse-graining a continuous state space,” *Entropy*, vol. 24, no. 10, p. 1432, 2022.
- [9] C. Tyler Diggans and Abd AlRahman R. AlMomani, *Boltzmann–Shannon interaction entropy: A normalized measure for continuous variables with an application as a subsample quality metric*, *Chaos*, Vol. 33, No. 12, p. 123131, 2023. DOI: 10.1063/5.0182349.
- [10] Landau, L. D., Lifshitz, E. M., *Statistical Physics*, Elsevier, 2013.
- [11] M. H. Partovi, *Entropic formulation of chaos for quantum dynamics*, *Phys. Lett. A* **151** (1990), no. 8, 423–428.
- [12] Roger Penrose, *The Road to Reality: A Complete Guide to the Laws of the Universe*, Jonathan Cape, London, 2005. ISBN: 9780224044479.
- [13] Arieh Ben-Naim, *Entropy Demystified: The Second Law Reduced to Plain Common Sense*, World Scientific Publishing Company, Singapore, 2007. ISBN: 978-9812700520.
- [14] Sethna, J. P., *Statistical Mechanics: Entropy, Order Parameters, and Complexity*, Oxford University Press, 2006.
- [15] Sinai, Ya. G., *On the Notion of Entropy of a Dynamical System*, *Doklady Akademii Nauk SSSR*, Vol. 124, pp. 768–771, 1959. (In Russian)
- [16] Lin, P., Bekey, G. A., and Abney, K., *Robots in War: Issues of Risk and Ethics*, 2009.

- [17] Xia, S., Zhang, C., Gao, W., Lv, F., and Wang, G., *Granular ball computing classifiers for efficient, scalable and robust learning*, Information Sciences, Vol. 621, pp. 61–74, 2023.
- [18] Yao, Y. Y., *Information granulation and granular relationships*, Proceedings of the IEEE International Conference on Granular Computing, pp. 326–329, 2008.
- [19] Zadeh, L. A., *Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic*, Fuzzy Sets and Systems, Vol. 90, No. 2, pp. 111–127, 1997.
- [20] Xia, S., Wang, Z., Zhu, X., et al., *Granular-ball computing: An efficient, robust, and interpretable adaptive multi-granularity representation and computation method*, International Journal of Approximate Reasoning, Elsevier, 2023.
- [21] Xia, S., et al., *Granular-ball computing and its applications in machine learning*, Information Sciences, Elsevier, 2024.
- [22] Diggans, Jordan et al., *Geometric Partition Entropy: A Framework for Entropy Estimation in Complex Systems*, Entropy, Vol. 24, MDPI, 2022.
- [23] Chen, Yang et al., *A Survey of Quantile Computation Algorithms*, IEEE Access, Vol. 8, IEEE, 2020.
- [24] Sparavigna, A. C., *Conditional Kaniadakis Entropy: A Preliminary Discussion*, 2015.
- [25] Khanam, S., Khanam, S., and Jaward, M., *Towards an Effective Intrusion Detection Model Using Focal Loss Variational Autoencoder for Internet of Things (IoT)*, Sensors, 22(15), 5822, 2022.

- [26] Pernice, R., Antonacci, Y., Zanetti, M., Busacca, A., Marinazzo, D., Faes, L., and Nollo, G., *Multivariate Correlation Measures Reveal Structure and Strength of Brain–Body Physiological Networks at Rest and During Mental Stress*, Frontiers in Neuroscience, 2021.
- [27] Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Physical Review E*, 69(6), 066138.
- [28] Kolmogorov, A. N., *A New Metric Invariant of Transitive Dynamical Systems and Automorphisms of Lebesgue Spaces*, Doklady Akademii Nauk SSSR, 119, pp. 861–864, 1958.
- [29] Luo, G., Chen, K., Tian, X., Qi, Y., & Zhang, Q. (2020). A survey of approximate quantile computation on large-scale data. *IEEE Access*, 8, 30575–30590.
- [30] Diggans, C. Tyler and AlMomani, Abd AlRahman R., *Generalizing Geometric Partition Entropy for the Estimation of Mutual Information in the Presence of Informative Outliers*, Chaos, vol. 35, no. 3, p. 033141, 2025, doi: 10.1063/5.0247397.
- [31] Cover, T. M. and Thomas, J. A., *Elements of Information Theory*, 2nd ed., Wiley-Interscience, Hoboken, NJ, 2006, ISBN: 978-0471241959.
- [32] Diggans, C. Tyler and AlMomani, Abd AlRahman R., *Geometric Partition Entropy: Coarse-Graining a Continuous State Space*, Entropy, 24(10), 1432, 2022.

APPENDIX A

MATLAB Implementation for High-Dimensional Mutual Information Estimation

The computational comparison between the partition intersection estimator and the granular ball estimator was implemented in MATLAB. The implementation used for synthetic data generation, estimation, and comparative visualization is provided in Listing [A.1](#).

```
1 function compare_PI_vs_GB_MI_GroundTruth()  
2  
3 clc;  
4 clear;  
5 close all;  
6 rng(1);  
7  
8 %  
=====
```

```

9  % Parameters
10 %
    =====
11
12 N      = 3000;
13 dims   = 2:2:10
14 ;
15
16 rho = 0.7;    % correlation coefficient
17
18 mi_true = zeros(size(dims));
19 mi_pi    = zeros(size(dims));
20 mi_gb    = zeros(size(dims));
21
22 %
    =====
23 % Main Loop
24 %
    =====
25
26 for i = 1:length(dims)
27
28     d = dims(i);
29
30     %
        -----

```

```

31     % Generate Gaussian-correlated data
32     %
33     %  $X \sim N(0, I)$ 
34     %  $Y = \rho X + \sqrt{1-\rho^2} Z$ 
35     %
36     %
37     -----
38
39
40     X = randn(N,d);
41
42     Z = randn(N,d);
43
44     Y = rho*X + sqrt(1-rho^2)*Z;
45
46     %
47     -----
48
49     % Exact Gaussian Ground Truth
50     %
51     %  $I(X;Y) = -(d/2) \log(1-\rho^2)$ 
52     %
53     %
54     -----
55
56     mi_true(i) = -0.5 * d * log(1-rho^2);
57
58     %
59     -----

```

```

54     % PI Estimator
55     %
56
57     mi_pi(i) = MI_PI(X, Y, 10);
58
59     %
60
61     % Granular Ball Estimator
62
63     mi_gb(i) = MI_GB(X, Y, 50, 30);
64
65     fprintf(['d=%2d|_TRUE=%.4f|_PI=%.4f|_GB=%.4f\n'],
66         ...
67         d, mi_true(i), mi_pi(i), mi_gb(i));
68 end
69
70 %
71 % Plot
72 %

```

73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97

```
=====

figure;

plot(dims, mi_true, '-k', 'LineWidth', 3);
hold on;

plot(dims, mi_pi, '-o', 'LineWidth', 2);

plot(dims, mi_gb, '-s', 'LineWidth', 2);

xlabel('Dimension_d');

ylabel('Mutual_Information');

title('PI_vs_GB_vs_Exact_Gaussian_Ground_Truth');

legend('Ground_Truth', 'PI_Estimator', 'GB_Estimator');

grid on;

end

%

% PI ESTIMATOR

%
```

=====

```

98
99 function I = MI_PI(X, Y, k)
100
101 N = size(X,1);
102
103 d = size(X,2);
104
105 %
106 % Quantile edges
107 %
108
109 edgesX = cell(d,1);
110 edgesY = cell(d,1);
111
112 for i = 1:d
113
114     edgesX{i} = quantile(X(:,i), linspace(0,1,k+1));
115
116     edgesY{i} = quantile(Y(:,i), linspace(0,1,k+1));
117
118 end
119
120 %

```

```

121 % Assign bins
122 %
123
124 binX = zeros(N,d);
125 binY = zeros(N,d);
126
127 for i = 1:d
128
129     [~,~,binX(:,i)] = histcounts(X(:,i), edgesX{i});
130
131     [~,~,binY(:,i)] = histcounts(Y(:,i), edgesY{i});
132
133 end
134
135 %
136 % Joint bins
137 %
138
139 joint_bins = zeros(N,1);
140
141 for i = 1:d
142     joint_bins = joint_bins * k + (binX(:,i)-1);
143 end
144

```

```

145 for i = 1:d
146     joint_bins = joint_bins * k + (binY(:,i)-1);
147 end
148
149 %
150 % Joint probabilities
151 %
152
153 [~,~,idx] = unique(joint_bins);
154
155 counts = accumarray(idx,1);
156
157 p_xy = counts / N;
158
159 %
160 % Marginal probabilities
161 %
162
163 px = marginal_prob(binX, k);
164
165 py = marginal_prob(binY, k);
166

```

```

167 %
168 % Entropies
169 %
170
171 Hx = -sum(px .* log(px + eps));
172
173 Hy = -sum(py .* log(py + eps));
174
175 Hxy = -sum(p_xy .* log(p_xy + eps));
176
177 I = Hx + Hy - Hxy;
178
179 end
180
181 %
182 % MARGINAL PROBABILITY
183 %
184
185 function p = marginal_prob(binMat, k)
186
187 N = size(binMat,1);
188

```

```

189 d = size(binMat,2);
190
191 bins = zeros(N,1);
192
193 for i = 1:d
194
195     bins = bins * k + (binMat(:,i)-1);
196
197 end
198
199 [~,~,idx] = unique(bins);
200
201 counts = accumarray(idx,1);
202
203 p = counts / N;
204
205 end
206
207 %
208     =====
209
210 % GRANULAR BALL MI ESTIMATOR
211 %
212     =====
213
214
215 function I = MI_GB(X, Y, maxBalls, minPts)
216
217 N = size(X,1);

```

```

215 data = [X Y];
216
217 %
    -----
218 % Initial Ball
219 %
    -----
220
221 balls = struct('idx', {1:N});
222
223 finalBalls = [];
224
225 %
    -----
226 % Recursive Splitting
227 %
    -----
228
229 while ~isempty(balls)
230
231     b = balls(1);
232
233     balls(1) = [];
234
235     idx = b.idx;
236

```

```

237     if length(idx) < minPts || length(finalBalls) >
238         maxBalls
239
240         finalBalls = [finalBalls b];
241
242         continue;
243
244     end
245
246     subData = data(idx,:);
247
248     [labels,~] = kmeans(subData,2,...
249         'MaxIter',50,...
250         'Replicates',2);
251
252     idx1 = idx(labels==1);
253
254     idx2 = idx(labels==2);
255
256     if isempty(idx1) || isempty(idx2)
257
258         finalBalls = [finalBalls b];
259
260     else
261
262         balls = [balls ...
263             struct('idx',idx1) ...
264             struct('idx',idx2)];
265
266     end

```

```

266
267 end
268
269 %
    -----
270 % Compute MI
271 %
    -----
272
273 I = 0;
274
275 for i = 1:length(finalBalls)
276
277     idx = finalBalls(i).idx;
278
279     nj = length(idx);
280
281     if nj < 5
282         continue;
283     end
284
285     pj = nj / N;
286
287     Xb = X(idx,:);
288     Yb = Y(idx,:);
289
290     % Covariance matrices
291

```

```

292     Cx = cov(Xb) + 1e-6*eye(size(X,2));
293
294     Cy = cov(Yb) + 1e-6*eye(size(Y,2));
295
296     Cxy = cov([Xb Yb]) + ...
297             1e-6*eye(size(X,2)+size(Y,2));
298
299     % Gaussian entropy approximation
300
301     Hx = 0.5 * log(det(Cx));
302
303     Hy = 0.5 * log(det(Cy));
304
305     Hxy = 0.5 * log(det(Cxy));
306
307     I = I + pj * (Hx + Hy - Hxy);
308
309 end
310
311 end

```

Listing A.1: MATLAB implementation for comparison of PI and GB mutual information estimators