

Pattern Discovery by Residual Analysis and Recursive Partitioning

Tom Chau, *Member, IEEE*, and Andrew K.C. Wong, *Member, IEEE*

Abstract—In this paper, a novel method of pattern discovery is proposed. It is based on the theoretical formulation of a contingency table of events. Using residual analysis and recursive partitioning, statistically significant events are identified in a data set. These events constitute the important information contained in the data set and are easily interpretable as simple rules, contour plots, or parallel axes plots. In addition, an informative probabilistic description of the data is automatically furnished by the discovery process. Following a theoretical formulation, experiments with real and simulated data will demonstrate the ability to discover subtle patterns amid noise, the invariance to changes of scale, cluster detection, and discovery of multidimensional patterns. It is shown that the pattern discovery method offers the advantages of easy interpretation, rapid training, and tolerance to noncentralized noise.

Index Terms—Pattern discovery, residual analysis, recursive partitioning, events, contingency tables.

1 INTRODUCTION

1.1 Motivation for Pattern Discovery

PATTERNS occur in all types of real-life data. The analysis of economic, financial, biomedical, demographic, and diagnostic data usually reveals some type of organization that is different from uniform randomness. Many authors refer to this organization as the structure of the data (see for example, Scott [1]), or equivalently, the patterns in the data [2]. We will use these intuitive terms interchangeably. With an understanding of the data's organization we can perform useful operations in the space of interest. These operations include the prediction of variable values, classification of previously unobserved samples and assessing the likelihood of a specific event. Therefore, the discovery of patterns is an indispensable step in the understanding of a given data set.

1.2 Related Work

Due to noise, sparsity, irregular geometries, and high-dimensionality, the uncovering of patterns in a data set can be a monumental task. Consequently, there is much varied literature and past research that is related to our work.

In the realm of pattern discovery in continuous data, existing approaches can be broadly lumped into four categories: kernel-based methods, multilayer perceptrons, exploratory techniques, and clustering-type procedures. These categories are not necessarily disjoint and multiple techniques could very well be applied to a given data set. In pattern recognition, the search for structure is implicitly tied to the problem of classification. In this paper, however, discrimination ability will not be the driving force of

discovery, although the results of discovery may be used for classification. We briefly overview each of the above broad class of approaches.

The first category encompasses a large body of literature, ranging from the classical work of Rosenblatt [3] and Parzen [4] to the more recent ideas of localized radial basis functions [5]. In general, these approaches seek a nonparametric description of the data's organization by a weighted summation of locally tuned basis functions. A common advantage of these methods is that with a proper choice of kernel, they provide a smooth density estimate over the space of interest. Contours of the smooth density estimate can reveal valuable information about the structure of the data. *Traditional kernel methods are supported by favorable asymptotic properties* (see [6], pp. 91-92), along with the general kernel theorem [1]. Radial basis functions have been shown to possess a universal approximation ability [7], [8]. Thus, theoretically, given unlimited numbers of hidden units and infinite samples, any pattern governed by a functional relationship can be detected and learned by the network. On the down side, in the absence of astronomical sample sizes, kernel-based nonparametric methods cannot simultaneously uncover both local and global organization in the data [1]. Radial basis function networks suffer from the curse of dimensionality when the intrinsic dimensionality of the data is less than the apparent dimensionality [9].

Multilayer perceptrons form an internal representation of the data's structure by an iterative training procedure such as back-propagation [10]. In terms of discovery, these networks are also universal approximators of continuous functions [11]. Further, they are most touted for their ability to internally represent arbitrarily complicated and previously unknown structure in high-dimensions amid substantial noise corruption. As pattern discovery mechanisms, some issues come to mind. First of all, perceptron-based neural methods are limited to learning many-to-one functional relationships and the learned representation is only a mean relationship [12]. This becomes a limitation when we consider data which may be multimodal in a

• T. Chau is with the Bloorview MacMillan Centre, 350 Runsey Rd., M4G 1R8, Toronto, Ontario, Canada.
E-mail: ortctc@oise.utoronto.ca.

• A.K.C. Wong is with Systems Design Engineering, University of Waterloo, N2L 3G1, Waterloo, Ontario, Canada.

Manuscript received 9 Apr. 1997; revised 17 Aug. 1999.

For information on obtaining reprints of this article, please send e-mail to: tkdc@computer.org, and reference IEEECS Log Number 104838.

one-to-many fashion. Secondly, the discovered dependencies and relationships are encoded incomprehensibly as weights in the trained network. Hinton diagrams [13] and bond diagrams [14] reveal limited insights into the learned representations. Recent work on rule extraction from trained neural networks [15], [16] strives to shed additional light on the discovered relationships (see [17] for a review). Although the correctness and completeness of the extracted rules can be difficult to verify, results reported in recent literature favorably support this type of discovery process.

Exploratory data analysis embodies a diverse collection of methods. Graphical methods attempt to pictorially illuminate the inherent structure in the data by the use of multiple scatter plots, glyphs, color, and parallel axes plots (see [1] for a review). These visual discovery techniques rely entirely on human interpretive abilities, which are no doubt powerful, but not immune to fatigue, error, and information overload. Projection methods such as principal component analysis (PCA) and projection pursuit (PP) are effective discovery tools when the relationships in the data are linear. When the structure lies on a nonlinear manifold, both methods have difficulty detecting the data's organization [18], [2].

Clustering methods seek out a special type of structure, namely, grouping tendencies in the data. In this regard, they are not as general as the other approaches but can provide valuable information when local aggregation of the data is suspected. The self-organizing map [19] is a special type of clustering algorithm [20] with greater discovery powers than traditional clustering methods. Although popular for its visual interpretability, the self-organizing map imposes a topological structure on the data. The assumed spatial topology may not be appropriate for all types of data [9].

Although no one method will suffice for all kinds of data, the aforementioned discovery techniques do share a common vision of excavating important information from a data set.

1.3 A Novel Pattern Discovery Method

In designing a pattern discovery system, some specific requirements come to mind. The discovery process and the representation of the patterns should be transparent and easily interpretable. To account for probabilistic uncertainty in the data, the description of patterns should have a probabilistic indication of statistical relevance. As real data is often noisy, the discovery system must be able to filter random noise from useful information. To alleviate an often vexatious data preprocessing step, the discovery process should not be sensitive to monotonic, linear scaling of the data.

In their attempts to meet these objectives, past methods have often sacrificed interpretability. In this paper, we present a pattern discovery method that aims at meeting the above challenges under a single framework, while remaining fairly transparent. The discovery method is based on the statistically guided construction of events in the sample space. In turn, event construction draws upon residual analysis in contingency tables and recursive partitioning of the sample space. The proposed approach is able to discover nonlinear and multimodal structure which may

be overlooked by standard exploratory techniques. It is also generic enough to detect functional relationships as well as clustering tendencies when they exist.

1.4 Paper Outline

The remainder of this paper is organized as follows. Section 2 lays down the fundamental definitions in the context of pattern discovery for continuous data. The discovery problem is formulated as a residual analysis problem in categorical data analysis. Section 3 presents the techniques of recursive partitioning and event merging for the purpose of pattern discovery. In Section 4, a probabilistic description of the discovered information is proposed. Section 5 summarizes the entire pattern discovery process with an implementable recursive algorithm. Experimental results are reported in Section 6, highlighting discovery capabilities and properties.

2 THEORETICAL BACKGROUND

The fundamental concepts and definitions are presented in two groups: those related to patterns and those related to discovery. Many definitions are adapted from probability theory and hypothesis testing.

2.1 Patterns

2.1.1 Feature

A feature, X , is a random variable which, in general, may be a nominal, ordinal, or numerical attribute or characteristic of an object or process. Our attention will be restricted to numeric features which are continuous-valued.

2.1.2 Event

An event is the elemental construct in the present formulation of pattern discovery. Our definition is analogous to the definition of events for random variables. When dealing with discrete data, the sample space is finite. Events are simply taken to be the elementary events [21], $X = x$, where X is the d -dimensional random vector and x is the vector of realizations [22].

In this paper, we are concerned with continuous data, particularly, the sample space \mathfrak{R}^d . A continuous sample space consists of an uncountable number of realizations. The use of elementary events becomes intractable, especially when defining probabilities [21]. Instead, events are defined as Borel sets.

Definition 1. Borel set [23].

Consider the sample space \mathfrak{R}^d . Let $-\infty < a_i < b_i < \infty$, $i = 1 \dots d$. Let $I_i = (a_i, b_i]$ be a one-dimensional semiclosed interval. A Borel subset of \mathfrak{R}^d is a d -dimensional rectangle, E , defined as:

$$E = I_1 \times I_2 \times \dots \times I_d = \{x | x_i \in I_i, 1 \leq i \leq d\} \quad (1)$$

where $x = [x_1 \dots x_d]^T$ is a point in \mathfrak{R}^d . The σ -field generated by the collection of all such rectangles is a Borel field of \mathfrak{R}^d , written as $B(\mathfrak{R}^d)$.

This leads directly to our definition of an event in the continuous sample space, \mathfrak{R}^d .

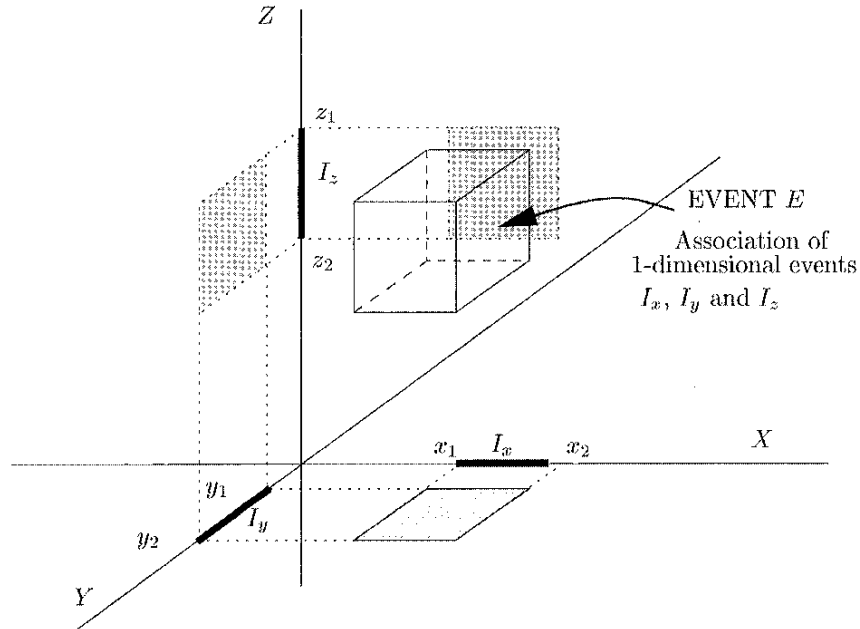


Fig. 1. An event in \mathbb{R}^3 .

Definition 2. Event.

An event in \mathbb{R}^d is a d -dimensional rectangle or Borel set.

Hence, an event is just a subspace of the continuous sample space. Fig. 1 portrays an event in \mathbb{R}^3 defined by the intervals I_x , I_y , and I_z . Here, X , Y , and Z represent three continuous-valued features. Two more definitions concerning events will be used frequently.

Definition 3. Volume.

Let E be a d -dimensional event defined by the intervals $\{I_i\}$, $i = 1 \dots d$. Let \mathcal{L}_i represent the length of the i th interval. The volume of E is given by,

$$v = \prod_{i=1}^d \mathcal{L}_i \tag{2}$$

Geometrically, this is just the volume of the d -dimensional rectangle, E .

Let $\{x_i\}$, $i = 1 \dots N$ represent the given set of N sample points within the sample space of interest.

Definition 4. Observed Frequency.

The observed frequency, n , of an event, E , is the number of sample data points, x_i , that fall inside the volume occupied by E .

2.1.3 Event Association

A d -dimensional event, E , can be interpreted as the association of lower dimensional events. For example, in Fig. 1, the event E can be viewed as an association of three one-dimensional intervals, I_x , I_y , and I_z . These one-dimensional intervals are themselves Borel sets on \mathbb{R} and, therefore, valid events. We summarize with a definition.

Definition 5. Event Association.

An event association is a joint occurrence of low-dimensional events. In particular, any d -dimensional event ($d \geq 2$), E , can be considered an event association, composed of d one-dimensional events (Equation (1)).

2.1.4 Pattern

In the context of this work, a pattern is simply a significant event or equivalently, if $d \geq 2$, a significant event association. Here, the term "significant" is used in the statistical context of hypothesis testing and will be clarified shortly.

2.2 Discovery

The discovery problem in a continuous sample space is to search for regions of organization, based on the information provided by the samples, $\{x_i\}$, $i = 1 \dots N$, drawn from an unknown population.

To elucidate on the discovery concepts, we first formulate the discovery problem in terms of categorical data analysis. This will allow the use of established and powerful statistical testing tools.

2.2.1 Discovery as Residual Analysis

Suppose that the continuous subspace of interest, $\Omega \subset \mathbb{R}^d$, is partitioned into J nonoverlapping regions. By the above definitions, we know that each of these subspaces is an event and, therefore, we label them as E_j , $j = 1 \dots J$. We are interested in discovering which events contain organized information. Equivalently, we want to discover the events which contain information that is not random. This latter viewpoint allows us to formulate the discovery problem as the analysis of a $2 \times J$ contingency table, under the assumption of product-multinomial sampling.

TABLE 1
 $2 \times J$ Contingency Table for Pattern Discovery

Population		Events				Totals
		E_1	E_2	...	E_J	
Actual data	1 (unknown distribution)	n_{11}	n_{12}	...	n_{1J}	n_{1+}
Virtual data	2 (uniform distribution)	n_{21}	n_{22}	...	n_{2J}	n_{2+}
Totals		n_{+1}	n_{+2}	...	n_{+J}	n_{++}

Consider the $2 \times J$ contingency table shown in Table 1. The columns of the table are labeled with the events, E_j , $j = 1 \dots J$. The first population represents the unknown distribution of the sample under consideration. Across this first row, we enter the actual observed frequencies, n_{1j} , $j = 1 \dots J$ of each event, E_j , $j = 1 \dots J$. The last column in the right is the row total, n_{1+} , where the + sign indicates summation over j . This notation is consistent with that of standard contingency table analysis [24], [25], [26], [27].

Since we desire to identify events which differ from randomness, we construct a second population that is governed by a uniform random distribution. The virtual frequencies across this row are calculated by,

$$n_{2j} = \frac{v_j}{V_{TOT}} n_{1+} \quad (3)$$

where v_j is the hypervolume occupied by event E_j , and

$$V_{TOT} = \sum_{j=1}^J v_j$$

is the total hypervolume of the subspace under consideration. As introduced earlier, n_{1+} is the actual number of true observations in the sample. Equation (3) assigns the virtual frequencies according to the fraction of the total volume occupied by each event. This is in accordance with our intuitive understanding of a uniform distribution. Note that the sum across the second row, n_{2+} , is equal to n_{1+} by construction. The bottom row of totals is the sum of the actual and virtual frequencies for each event.

We now have a $2 \times J$ table, consisting of two independent multinomial populations. This is a case of product-multinomial sampling. We would like to compare the two populations for significant local differences. In contingency table parlance, this amounts to a test for homogeneity of proportions. The null hypothesis, H_0 , is given by,

$$H_0 : p_{1j} = p_{2j}, \quad j = 1 \dots J \quad (4)$$

where p_{ij} is the probability of event j for population i .

To detect local departures from this hypothesis, we require estimates of the expected values, m_{ij} , under the assumption that H_0 is true. We now draw upon a number of standard results for a two-dimensional table with product-multinomial sampling. Since each n_{ij} has a multinomial

distribution, the expected value is, $m_{ij} = n_{i+} p_{ij}$ [24], [28], [27]. Assuming the null hypothesis, H_0 , to be true, the estimate of the common value of p_{ij} is,

$$\hat{p}_{ij} = \frac{n_{+j}}{n_{++}} \quad (5)$$

From this we obtain as the estimated expected value,

$$\hat{m}_{ij} = n_{i+} \frac{n_{+j}}{n_{++}} \quad (6)$$

Using the definition $n_{+j} = n_{1j} + n_{2j}$, and exploiting the symmetry relations, $n_{1+} = n_{2+}$ and $n_{++} = 2n_{1+}$, we may specialize (6) to the present table. Doing so, we arrive at

$$\hat{m}_j = \frac{1}{2} n_{+j} \quad (7)$$

$$\hat{m}_j = \frac{1}{2} (n_{1j} + n_{2j}) \quad (8)$$

where n_{2j} is given by (3). Equation (8) is the estimated expected value for the cells in j th column of the contingency table, assuming H_0 is true. Note that we've dropped the i subscript, as the expected value \hat{m} is independent of the population i , by virtue of the hypothesis of homogeneity.

Since we are interested in detecting deviations from H_0 at the level of individual cells, we invoke the adjusted residual test statistic [29], rather than a global Pearson chi-square statistic. The adjusted residual is an asymptotically standard normal test statistic [29], [24]. For the cell corresponding to the i th population and j th event, it is defined by [29],

$$r_{ij} = \frac{n_{ij} - \hat{m}_{ij}}{\hat{c}_{ij}^{1/2}} \quad (9)$$

where $\hat{c}_{ij}^{1/2}$ is the estimated asymptotic variance of the numerator. For a two-dimensional table, under the assumptions of product-multinomial sampling and homogeneity of proportions, this variance is given by [24], [25],

$$\hat{c}_{ij} = \hat{m}_{ij} \left(1 - \frac{n_{+j}}{n_{++}}\right) \left(1 - \frac{n_{+i}}{n_{++}}\right) \quad (10)$$

By applying (7) and the relations, $n_{++} = 2n_{1+}$ and $n_{i+} = n_{1+}$, (10) is simplified to,

$$\hat{c}_j = \frac{1}{2} \hat{m}_j \left(1 - \frac{\hat{m}_j}{n_{1+}} \right) \tag{11}$$

where again the i subscript has been dropped as there is no dependence on the population i .

Recall that only the first row of the contingency table (Table 1) describes the actual data. Hence, we will only compute residuals for the *first* row of cells. For simplicity of notation, these residuals will be written as r_j without the population subscript. Substituting (8) into the numerator of the residual definition, (9), we arrive at the adjusted residual, tailored for our table,

$$r_j = \frac{\frac{1}{2}(n_{1j} - n_{2j})}{c_j^{1/2}} \tag{12}$$

with \hat{c}_j defined by (11). The above statistic is suitable for detecting differences between the event frequencies of the unknown distribution against those of a uniform random distribution. It has been shown that $r_j \rightarrow N(0, 1)$ in distribution if the assumptions ([27], p. 439) and regularity conditions [30], ([27], p. 426) used in the asymptotic derivation ([29], p. 211), ([24], p. 226), [30] hold.

We are now in a position to define a significant event. Let z_β be the value of the standard normal deviate, $Z \sim N(0, 1)$, such that $P(Z \leq z_\beta) = \beta$. Let \hat{r}_j be the estimated value of the adjusted residual. The following definitions are analogous to those in statistical hypothesis testing.

Definition 6. (Positively) Significant Event.

An event, E_j , is significant at the $\alpha \times 100$ percent significance level, if $\hat{r}_j \geq z_{1-\alpha/2}$.

For example, at the 5 percent significance level, an event E_j would be deemed significant if $\hat{r}_j \geq 1.96$. A significant event indicates a breach of the homogeneity hypothesis. Spatially, significant events are parts of the subspace that contain organized data. In this paper, we reserve the terminology, "significant event" for positively significant events, unless specified otherwise.

Definition 7. Negatively Significant Event.

An event, E_j , is negatively significant at the $\alpha \times 100$ percent significance level, if $\hat{r}_j \leq z_{\alpha/2}$.

Negatively significant events are common in practice. They indicate that the observed frequency of an event is much less than the corresponding virtual frequency of the uniform random distribution. In terms of geometry, these events delineate regions of the subspace, Ω , where the data is organized, but notably sparse.

For the sake of thoroughness, we also give the definition of insignificant events.

Definition 8. Insignificant Event.

An event, E_j , is insignificant at the $\alpha \times 100$ percent significance level, if $|\hat{r}_j| < z_{1-\alpha/2}$.

Insignificant events correspond to events which do not violate the homogeneity assumption at the chosen level of significance. Geometrically, these events demarcate sections of the subspace, Ω , where the data is uniformly random, or lacking in organization.

In this section, we have explained how to construct a $2 \times J$ contingency table for pattern discovery. Computing residuals for the cells in the first row of this table, we detect departures from the homogeneity hypothesis. By so doing, we are actually identifying events with observed frequencies not governed by the uniform random distribution. In turn, these events correspond to regions of the original sample space that contain some sort of organized information. In this way, the problem of discovering informative subspaces is posed as a residual analysis problem of detecting departures from the homogeneity of proportions assumption.

2.2.2 Pattern Discovery

In light of the above definitions, pattern discovery is the search for significant events or significant associations among events. Hence, pattern discovery is a mathematical optimization problem where our objective is to size and locate events to maximize the adjusted residual statistic (12). Unfortunately, the residual statistic is a nonsmooth objective and gradients cannot be computed. Further, the statistic is ill-posed in that cell frequencies cannot be analytically expressed in terms of the event size and location. In the next section of this paper, we discuss a method of obtaining an approximate solution to this optimization problem.

3 PARTITIONING

Given a data set, $\{x_i, i = 1 \dots N\}$, and a set of events, $\{E_j\}$, the above discussion describes how to assess the significance of each event. We now address the problem of practically constructing events in the given sample space. The following scheme is a recursive approximation in lieu of mathematical optimization.

3.1 Partitioning a Subspace $\Omega \subset \mathbb{R}^d$: Theoretical Justification

Our approach to event construction is to partition the continuous sample space into nonoverlapping sections. According to the above definitions, we know that such subspaces constitute valid events. However, in practice, we limit our attention to a subspace Ω of \mathbb{R}^d . Fortunately, partitioning this subspace alone is sufficient for the generation of a Borel field of \mathbb{R}^d (see Definition 1). Consider \mathbb{R}^d as the union of the subspace Ω and its complement Ω^C in \mathbb{R}^d , i.e., $\mathbb{R}^d = \Omega \cup \Omega^C$. We can consider Ω^C to be arbitrarily partitioned into countably many nonoverlapping rectangles. In this way, a partition of the subspace Ω always provides a partition for the sample space \mathbb{R}^d (see Fig. 2). This abstract argument will be applied in building a probabilistic description of the data. For now, we may be content that restricting our attention to a subspace, Ω , will not violate the aforementioned Definitions 1 and 2.

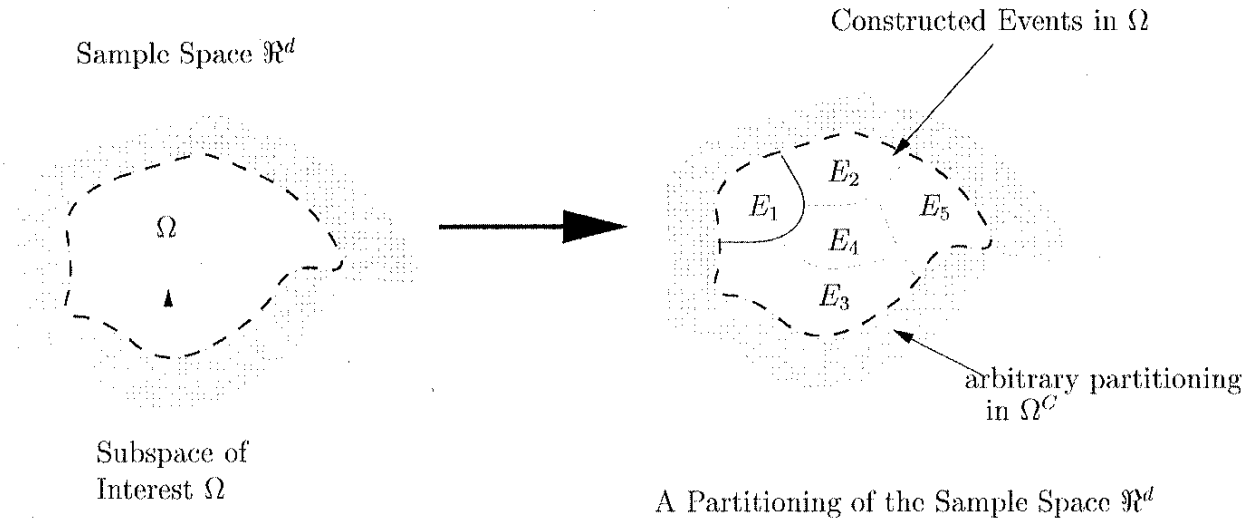


Fig. 2. Partitioning the sample space.

3.2 Marginal Maximum Entropy Partitioning

In accordance with Definition 1, we must partition Ω into countably many nonoverlapping rectangular regions. Clearly such a partition is not unique. However, since we eventually want to describe the data's organization probabilistically, we desire a technique which maximally preserves the data's probability distribution. One such technique is known as marginal maximum entropy partitioning (MMEP) [31], [32]. This method segments the subspace of interest while approximately maximizing the overall entropy [33], H , of the partition.

$$H = - \sum_{j=1}^J P(E_j) \log P(E_j) \quad (13)$$

where $P(E_j)$ is the probability associated with event E_j and the summation is over all events. The approximate nature of the method is due to the fact that partitioning is not performed in the full dimensionality, but rather *marginally* along each dimension. Lascurain [32] has shown that this is a reasonable approximation [32] for a number of distributions.

It is well known that the entropy of a partition (13) is maximized when all probabilities are equalized [33], [34]. In the present application, event probabilities can be estimated intuitively by the proportion of samples contained in each event. If we let N be the total number of samples under consideration, then the estimated probability of event E_j is

$$\hat{P}_j = n_j / N \quad (14)$$

Consequently, the equalization of probabilities translates into an equalization of event frequencies. Therefore, marginal maximization of entropy involves the segmentation of individual axes into one-dimensional intervals such that each contains an approximately equal number of samples. We summarize the procedure of Lascurain [32] below.

3.2.1 Marginal Maximum Entropy Partitioning Procedure (MMEP)

Let \mathcal{E} represent the subspace to be partitioned. The partition size will be denoted as Q . As usual, d is the dimension of the data. Partitioning will produce $Q \times Q \times \dots \times Q = Q^d$ events. The operation $\lfloor \cdot \rfloor$ rounds down to the nearest integer:

1. Enumerate the number of sample points, N , in \mathcal{E} . Set $i = 1$.
2. For axis i ,
 - a. Identify the minimum and maximum value of the i th coordinate. Label them as a_1 and a_{Q+1} , respectively.
 - b. Choose $Q - 1$ points, a_2, \dots, a_Q along the axis, between a_1 and a_{Q+1} , so that there are Q intervals, each containing $\lfloor N/Q \rfloor \pm 1$ points.
 - c. From each a_i , extend a $(d - 1)$ dimensional plane perpendicular to the axis.
 - d. If $i < d$, increment i and return to step 2. Otherwise, go to step 3.
3. The intersection points of the $(Q + 1)^d$ planes define Q^d events in \mathbb{R}^d .

For ease of visualization, the procedure is illustrated in Fig. 3 with a two-dimensional example ($d = 2$), using a partition size of $Q = 3$, for a total of $Q^d = 3^2$ events. The data consists of $N = 45$ points. The dashed lines indicate the locations of the partition points on each axis. The vertical and horizontal partitions are shown separately on the left side. They are combined to produce the right-hand pictorial. Each interval contains $N/Q = 15$ points and the observed frequencies of each event are approximately uniform.

The choice of hypercube partitions is not unique. The hypercube is, however, the simplest geometry which satisfies the definition of an event, requiring only 2^d parameters for complete specification. To specify a hyper-ellipsoid, for instance, one would need in the order of d^2

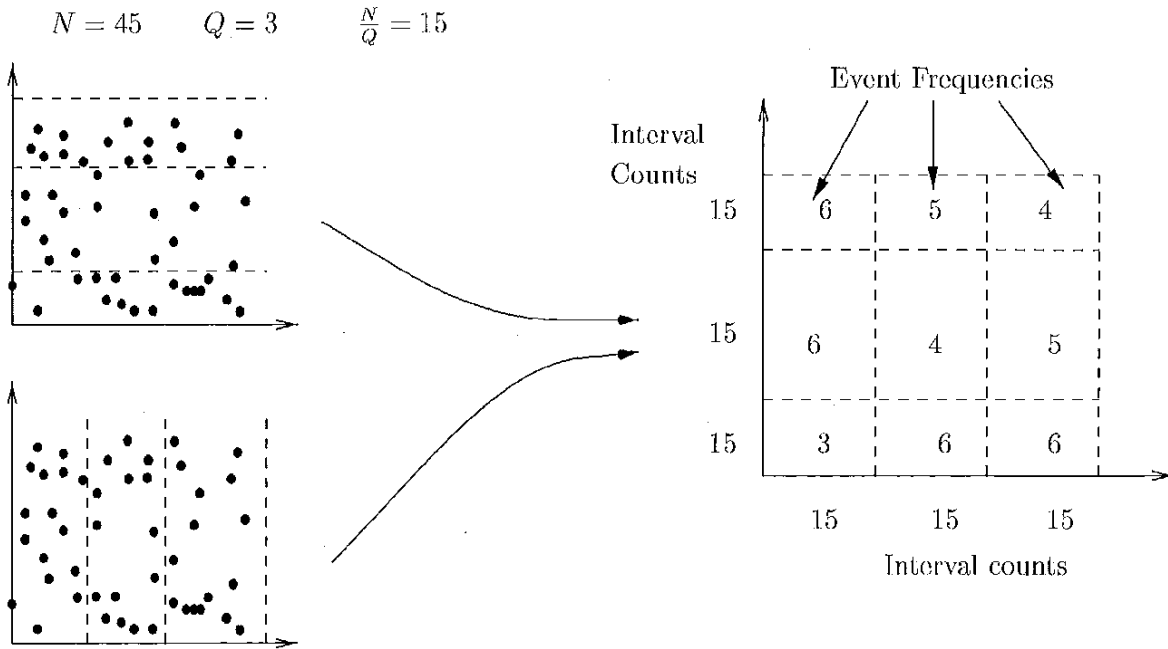


Fig. 3. Example of MMEP with two-dimensional data.

parameters. Any loss in representational accuracy resulting from the use of simple hypercubes is mitigated by recursive partitioning, described in Section 3.4.

3.3 Boundary Refinement

The MMEP method is fairly rigid in that the boundaries of the constructed events must lie on the partitioning planes. For discovery purposes, this is inadequate, as data on curved surfaces will not be well-represented. The left panel of Fig. 4 exemplifies this dilemma with four events, $\{E_1, \dots, E_4\}$. A simple enhancement is to adjust the event boundaries to coincide with the maximum and minimum coordinates of the contained data. The adjacent pictorial demonstrates this adjustment. In this way, the location of events can be completely general. Close adherence to the actual data is ensured, despite the restrictive nature of the partitioning scheme. These refined events undoubtedly satisfy the proposed theoretical framework.

3.4 Recursive Partitioning

From the above example, it is evident that the constructed events are rather coarse in that they do not

effectively capture the organization of the data. To refine the events, we need to extend the above partitioning procedure in a recursive manner. In the present context, recursive partitioning [31], [35], [36], [37], [38], [39], [40] means that we again apply the MMEP procedure to each constructed event. Clearly, it is not meaningful to continue this process indefinitely. Termination conditions are required (see Section 3.6). Moreover, recursive partitioning of every constructed event may not be necessary, especially where there is actually little data or information. Some criterion is needed to guide the process of *selective* recursive partitioning [31], [35].

Unlike previous schemes based upon event covering criteria [31], [35], [41], we use the criterion of significance introduced in Section 2.2.1 to direct the path of recursive partitioning. Plainly put, only events which are significant become candidates for additional partitioning. Fig. 5. portrays an example of two levels of partitioning. The initial partition is shown on the left, along with the residual values of each event. On the right is the result of recursive partitioning. Here, we have used a 5 percent significance

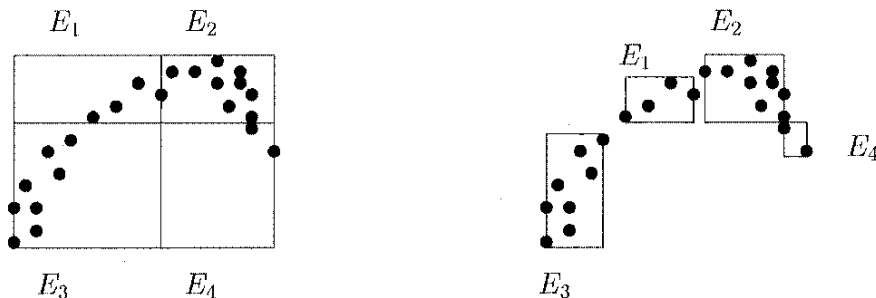


Fig. 4. Boundary refinement.

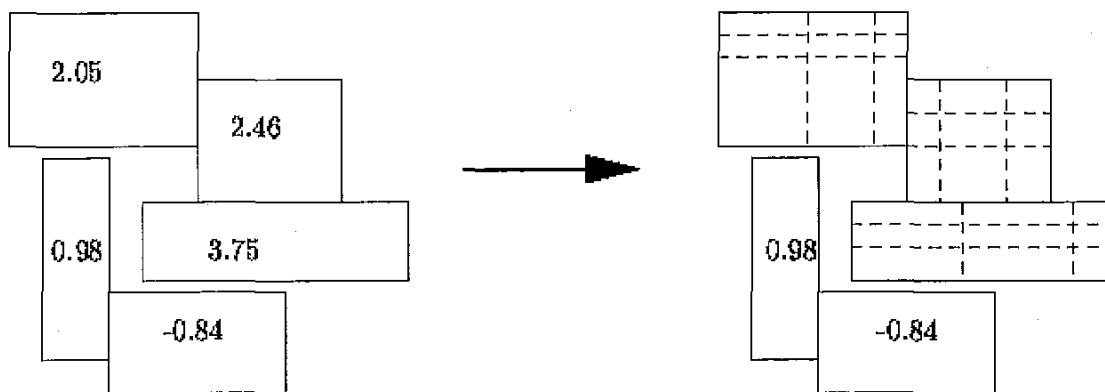


Fig. 5. Recursive partitioning.

level, so that only events with residuals larger than or equal to 1.96 are repartitioned.

3.5 Adaptive Partition Size

In the original work on MMEP [32] and early recursive partitioning-based discovery schemes [31], [35], there is no mention of how to determine an appropriate partition size, Q .

Recall that our main objective is to discover organization in the data. Therefore, at each level of partitioning, it makes sense to choose a partition size which yields the most information about the data's structure. Since significant events (both positive and negative) capture the data's organization, we choose the partition size, Q , to maximize the ratio of significant events to all candidate events. Formally, suppose the sample space is to be partitioned into Q^d events. Let $M(Q)$ represent the number of positively

and negatively significant events in this partition. Choose Q such that,

$$\max \frac{M(Q)}{Q^d}. \tag{15}$$

This objective function encapsulates the trade-off between maximizing the number of significant events while minimizing Q , thereby minimizing computational complexity. Intuitively, we see that the value of (15) will increase as Q is incremented integrally from an initial value of unity. However, for large Q , (15) will decrease monotonically. It can be shown that (15) possesses exactly one maximum. Maximization of the above criterion is an integer program and can be solved by the standard branch and bound algorithm [42].

As a simple example, suppose that we were trying to detect the structure of the data in Fig. 6. The left panel is the

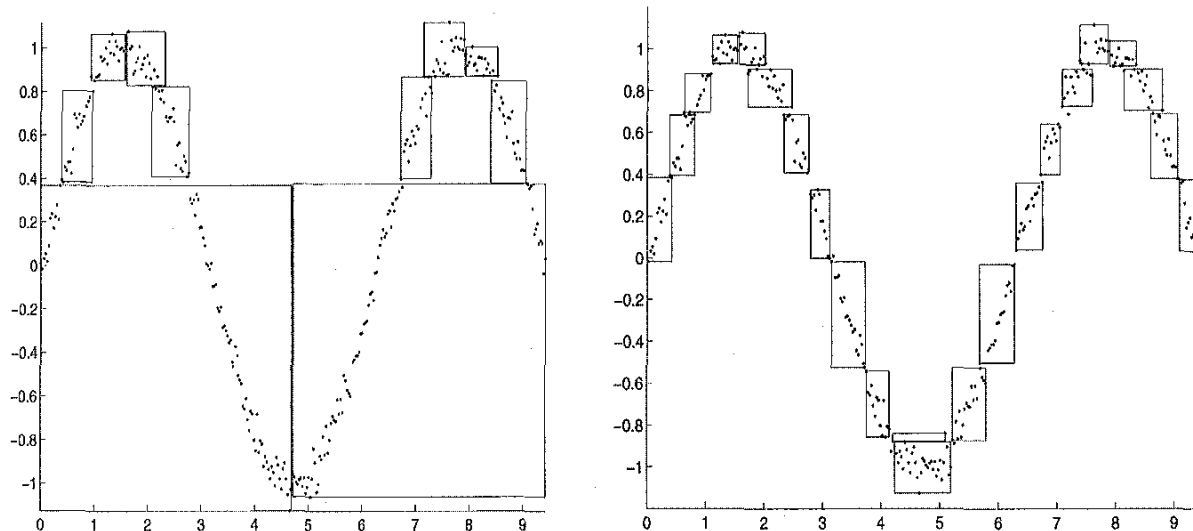


Fig. 6. Adaptive partitioning.

result of fixed 2×2 partitioning. Much of the structure in the data is lost. The right panel exemplifies the substantial improvement in discovery due to adaptive partitioning at each level of recursion.

3.6 Termination Conditions

As alluded to in the explanation of recursive partitioning, criteria need to be established to prevent infinite recursion. We now outline these conditions under which partitioning of an event will cease. More than one condition may be satisfied simultaneously:

1. $\Lambda > 20\%$

Here, Λ is the proportion of events with expected frequency below 5. The first condition states that partitioning should cease at a given level if more than one-fifth of the events have expected frequencies less than 5. This constraint, mentioned in [27], maintains the validity of statistical testing.

2. $v_j < \frac{V_{TOT}}{n_{i+}} (2\zeta_{min} - n_{i,j})$

This condition prevents an event, E_j , from being partitioned when the expected event frequency, $\hat{m}_{i,j}$, is below the minimum value, ζ_{min} , required for reliable statistical testing. The derivation of this condition is provided in Appendix A.

3. $\hat{r}_j \leq z_{\alpha/2}$

When an event is negatively significant, it should not be partitioned. In fact, subspaces with little or no data are removed from further consideration.

4. $|\hat{r}_j| < z_{1-\alpha/2}$

This last condition halts recursive partitioning when an event is insignificant. Since the event's frequency does not deviate from that of uniformity, there is no need for additional investigation of the subspace.

3.7 Event Merging

Although recursive partitioning permits the detection of local organization in the data, two unacceptable complications arise. The first issue is computational. By a straightforward calculation, we see that with increasing levels of recursion, the number of events grows exponentially. Apart from being an exhaustive drain on resources, an overwhelming number of events is unmanageable for tasks such as classification or prediction. Secondly, overlapping events would arise if we were to store every event generated from multiple levels of recursion in a given subspace. From a theoretical perspective, Definitions 1 and 2 prohibit overlapping events.

Fortunately, these problems can be effectively alleviated by *merging* insignificant events within the current subspace, \mathcal{E} , under consideration. Suppose the MMEP procedure is applied to a subspace \mathcal{E} , using a partition size Q , to produce a set of events $\{E_j\}, j = 1 \dots Q^d$, such that

$$\mathcal{E} = \bigcup_{j=1}^{Q^d} E_j.$$

An index set, κ , can be formed to identify all the insignificant events.

$$\kappa = \{j \mid |\hat{r}_j| < z_{1-\alpha/2}\} \tag{16}$$

where $1 \leq j \leq Q^d$. Once identified, these insignificant events, $\{E_j\}, j \in \kappa$, can be merged into a compound event, $E_c = \bigcup_{j \in \kappa} E_j$. The following is a synopsis of the procedure. Assume that the space \mathcal{E} has been partitioned by the MMEP procedure.

3.7.1 Event Merging Procedure

1. Identify insignificant events, $\{E_j\} \in \mathcal{E}, j \in \kappa$. The index set, κ , is defined by (16). If there are no insignificant events, then stop.
2. Compound event construction:
 - a. Compute the compound event volume, $v_c = \sum_{j \in \kappa} v_j$.
 - b. Compute the compound event frequency, $n_c = \sum_{j \in \kappa} n_j$.
 - c. Compute the compound event probability density value, $p_c = \frac{n_c}{N \cdot v_c}$, where N is the total number of samples under consideration.
3. Remove partition boundaries between insignificant events.

Probability density estimation as in step 2c will be discussed in Section 4. It is included here for completeness. Fig. 7 exemplifies the effect of event merging. The events and their residual values are shown on the left. Again, we use a 5 percent significance level to demarcate the significant events ($\hat{r}_j > 1.96$). On the right, the significant events are retained (shown as shaded rectangles) while the insignificant events ($|\hat{r}_j| < 1.96$) are merged.

With regard to the computational problem, event merging offers substantial relief. Suppose that we apply MMEP to \mathcal{E} , using a partition size of Q . Without event merging, MMEP will produce $Q^d = M + |\kappa|$ events, where M represents the number of positively and negatively significant events and $|\kappa|$ is the number of insignificant events. With merging, the tally of events reduces to $M + 1$. As the number of insignificant events typically far outweighs its significant counterpart, i.e., $|\kappa| \gg M$, merging amounts to an order of magnitude reduction in the total number of events stored.

From a theoretical standpoint, there will no longer be overlapping events. The compound events may assume quite arbitrary geometries but are still consistent with the event definitions. The argument is that the compound event is a union of a finite number of events and is therefore itself an event, by virtue of the closure property of $B(\mathfrak{R}^d)$.

4 PROBABILISTIC DESCRIPTION AND SCALE INVARIANCE

After recursive partitioning and residual analysis, the subspace is divided into significant (both positive and

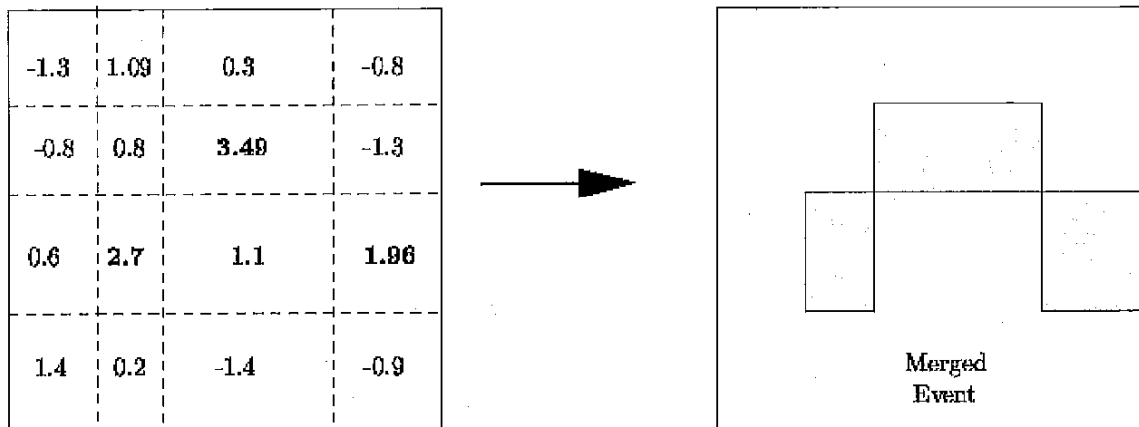


Fig. 7. Event merging.

negative) and insignificant events. At this point, a probabilistic description of the data's organization can be constructed. Such a description is useful for data interpretation, classification, and prediction. We propose a straightforward discrete density estimate as the appropriate probabilistic characterization of the data's organization.

4.1 Probabilistic Description

To simplify notation, we define the event indicator function.

Definition 6. *Event Indicator Function.*

The indicator function for the event, E_j , is defined as,

$$I_{(j)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x} \in E_j \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

Here, $\mathbf{x} \in \mathbb{R}^d$ is a point in the d -dimensional sample space.

Employing Definitions 3 and 4 for event volume and observed frequency, we can assign to an event, E_j , the following probability density estimate,

$$\hat{p}_j = \frac{n_j}{N \cdot v_j} \quad (18)$$

where, as usual, N is the total number of sample points under consideration. This definition is along the lines of the general nonparametric density estimate of Duda and Hart [6]. Note that this probability density is just the probability estimate of (14), \hat{P}_j , divided by the event volume, v_j . We see immediately that the normalization condition is satisfied:

$$\sum_j \hat{P}_j = \sum_j \hat{p}_j \cdot v_j = 1 \quad (19)$$

To obtain a discrete probability density function (PDF), \hat{f} , valid for the whole sample space, we recall that the events, E_j , do not overlap, and therefore we may write compactly,

$$\hat{f}(\mathbf{x}) = \sum_j I_j(\mathbf{x}) \hat{p}_j \quad (20)$$

where, again, $\mathbf{x} \in \mathbb{R}^d$ and $I_j(\mathbf{x})$ is the indicator function previously defined. Note that only one term in the summation will have $I_j(\mathbf{x}) \neq 0$ as the data point can only

fall into one event. Due to the frequency equalization tendency of the partitioning process, the above discrete density function is scale invariant.

4.2 Scale Invariance

We first clarify what we mean by monotonic scaling, a special case of monotone transformations of the coordinate axes ([43], pp. 22-24). Suppose we have a data set $\mathbf{X} = \{\mathbf{x}_i\}$, $i = 1, \dots, N$, and each $\mathbf{x}_i \in \mathbb{R}^d$.

Definition 10. *Monotonic Scaling.*

A data set $\mathbf{X} = \{\mathbf{x}_i\}$, $i = 1, \dots, N$ is said to undergo a monotonic change of scale to $\mathbf{X}' = \{\mathbf{x}'_i\}$, $i = 1, \dots, N$, if for each dimension, k , $k = 1, \dots, d$,

$$\mathbf{x}'_{ik} = \gamma_k \mathbf{x}_{ik} \quad \gamma_k > 0 \quad (21)$$

where γ_k are, in general, d distinct positive constants. In other words, the k th coordinate of every data point is scaled by the same positive factor, γ_k .

By scale invariance, we mean that the probability estimates are invariant to change of scale or equivalently, that the ratio of the densities of any two events is invariant. This concept is formalized below.

Suppose we have two data sets, $\mathbf{X} = \{\mathbf{x}_i\} \in \Omega$, and $\mathbf{X}' = \{\mathbf{x}'_i\} \in \Omega'$, $i = 1, \dots, N$. Let \mathbf{X}' be a monotonically scaled version of \mathbf{X} , as defined by (21). Suppose now we apply MMEP to Ω and Ω' . By nature of (15), the same partition size, Q , will be selected for both data sets. The result of partitioning will be two sets of events $\{E_j\}$ and $\{E'_j\}$, $j = 1, \dots, Q^d$. We have the following proposition.

Proposition 1. *Scale Invariance.*

Let $\{E_j\}$ and $\{E'_j\}$ be constructed as above. We will write the probability of an event, E_j , explicitly as $\hat{P}(E_j)$. Then,

$$\hat{P}(E_j) = \hat{P}(E'_j) \quad j = 1, \dots, Q^d \quad (22)$$

Furthermore, with $i \neq j$,

$$\frac{\hat{p}(E_i)}{\hat{p}(E_j)} = \frac{\hat{p}(E'_i)}{\hat{p}(E'_j)} \quad (23)$$

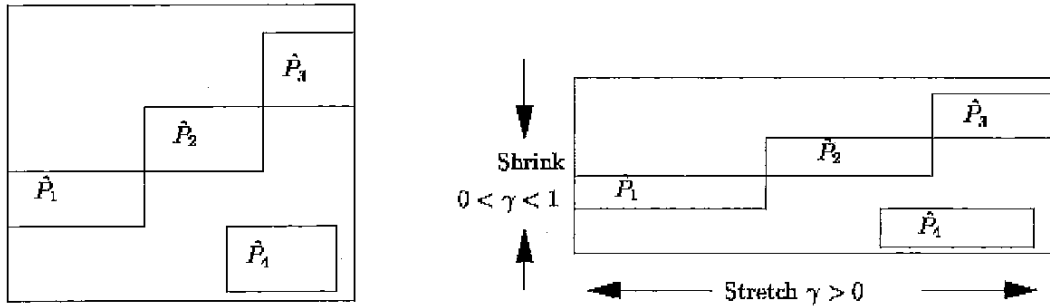


Fig. 8. Scale invariance: Estimated probabilities \hat{P}_j are unchanged.

where $\hat{p}(E_j)$ is the probability density value of event E_j as given by (18).

This property is a direct result of MMEP and the chosen density estimate. Verification is provided in Appendix B. A pedagogical example is provided in Fig. 8, where the scale of one feature is expanded and the other is compressed. The importance of scale invariance in the related area of nonparametric pattern classification was recognized by Devroye [44] and Friedman [38]. Numerous partitioning schemes were designed specifically with this property [39], [38], [45], [31], [35]. In discovery, this is an equally indispensable property. From a functional perspective, a discovery system should uncover the same patterns (events) from a data set, regardless of how the features may be scaled. Further, the probabilistic description of these patterns should be consistent and independent of scale.

In data preprocessing, the determination of a suitable normalization factor is not always straightforward and linear. Indeed, the process is often laborious and ad hoc. Scale invariant residual analysis offers a structured and objective discovery alternative.

5 PATTERN DISCOVERY ALGORITHM

We summarize the above discussions by presenting the pattern discovery algorithm based on residual analysis and recursive partitioning. The algorithm draws upon partitioning ideas from [39], [38], [32] and the hierarchical methodology of [31], [35].

5.1 Discovery Algorithm: Residual Analysis and Partitioning (RAP)

1. Suppose we want to seek patterns in a subspace Ω . Set $\mathcal{E} = \Omega$.
2. Compute the partition size Q for \mathcal{E} , according to the criterion of (15).
3. Partition \mathcal{E} into Q^d events, $\{E_j\}$, $j = 1 \dots Q^d$, using the MMEP approach of Section 3.2. Set $j = 1$.
4. For event E_j ,
 - a. Refine its boundaries by the method described in Section 3.3.
 - b. Compute its associated virtual frequency, \hat{n}_{2j} , according to (3).

- c. Compute the probability density estimate, \hat{p}_j , according to (18).
 - d. Compute the adjusted residual, \hat{r}_j , using (12).
5. Store, recurse, or continue:
 - a. If termination condition (4) is met, mark the event as insignificant. If termination condition (1) or (2) is satisfied, store this event as significant or mark it as insignificant, depending on the value of its residual. If condition (3) is true, simply remove this event from further consideration. Proceed to step 5c.
 - b. **Recursion.** If no termination conditions are met, set $\mathcal{E} = E_j$ and go to step 2.
 - c. Increment j and proceed to examine the next event by returning to step 4. If all events have already been examined, i.e., $j = Q^d$, then go to step 6.
6. Apply the Event Merging procedure of Section 3.7 to \mathcal{E} . If a compound event is constructed, store the compound event. If this is a nested recursion, return to step 5c. If this is the top-level of partitioning, stop.

It is worthwhile to note that, in theory, and in practice, events with 0 density value, $p_j = 0$, can be harmlessly discarded.

6 EXPERIMENTAL RESULTS

The experimental results are organized into two sections. The first part will elucidate basic properties of the pattern discovery technique while the second part illustrates the application to a multidimensional data set. Experiments with other methods are included only to facilitate the explanation of pattern discovery properties. These experiments are not intended to be an exhaustive comparison with existing approaches. Throughout the discussion, we will demonstrate the ease of interpreting the discovery results. For all experiments, we have used a 5 percent significance level to gauge significant events. Also, in all experiments, the time required for the discovery was negligible, never exceeding 10 seconds.

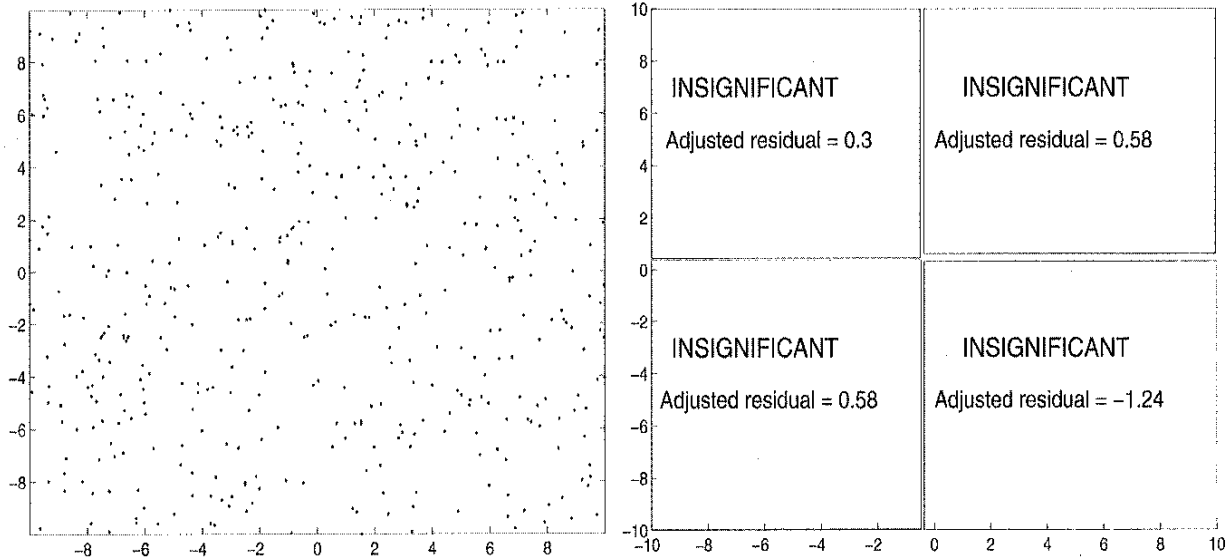


Fig. 9. Rejection of uniform randomness in discovery.

6.1 Basic Properties

6.1.1 Uniform Random Noise Rejection

To begin, we demonstrate the property of noise rejection when the data contains no organization. Consider the uniform random bivariate data (see Fig. 9). As shown on the right side, the system discovers no significant events. This is reassuring since we do not wish to discover structure which does not really exist in a data set. What happens if there *is* structure embedded in the noise? This leads to the next example.

6.1.2 Discovery Amid Noncentralized Noise

The second example examines the property of noise tolerance in the discovery of nonlinear structure. Fig. 10 shows the nonlinear relationship between intensity and mean angle taken from a physics problem. The governing equation is

$$y = k_1(1 + \cos(2x)^2) \frac{\exp(k_2 \sin(x))}{\sin(x)^2} + \delta \quad (24)$$

where $k_1 = 1$ and $k_2 = -1$ are constants and δ is a Gaussian variable with 0 mean and standard deviation 10^{-3} . From the governing relationship, 1,000 points were generated with additive Gaussian noise. In addition, the data has been corrupted with 250 uniform background noise points (x, y) with $0.89 \leq x \leq 2.24$ and $0.69 \leq y \leq 0.79$. Application of RAP resulted in the discovery of 34 events (Fig. 11). Notice that the noise has essentially been filtered out.

We can proceed to make use of the discovered information for predicting intensity values for given mean angles. Based on the data contained in each event, a centroid point can be computed. These centroids can then be spline-fitted. Although beyond the scope of this paper, it is worthwhile to mention that this spline fitting result could be used for

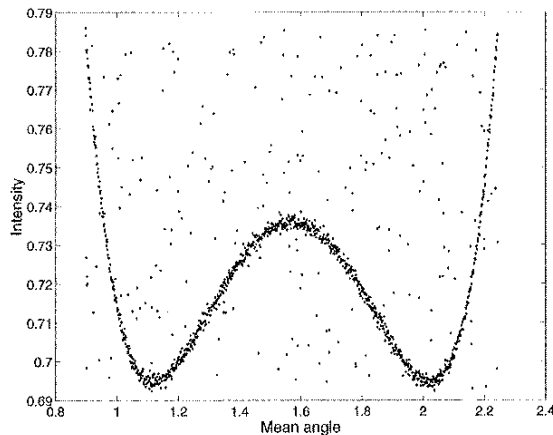


Fig. 10. Raw data.

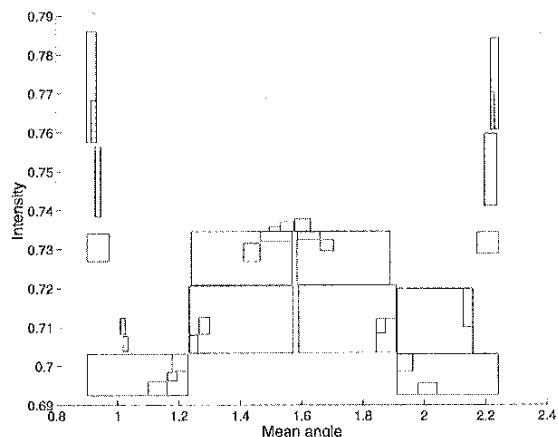


Fig. 11. Discovered events.

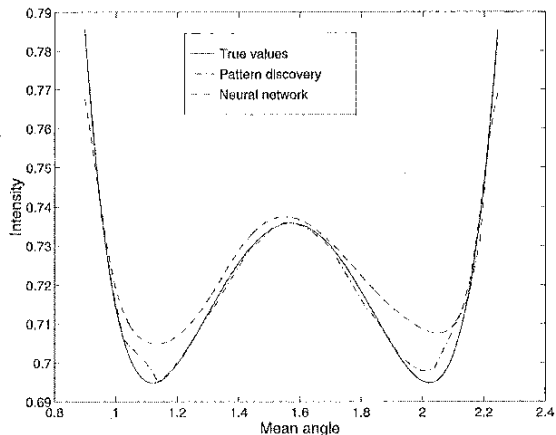


Fig. 12. Predicted values.

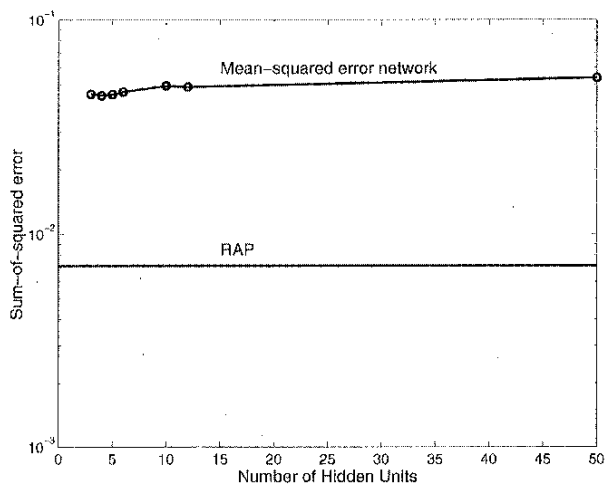


Fig. 13. Sum-of-squared errors.

scientific or functional form discovery or even for nonlinear regression.

We can predict intensity values for given mean angles, simply by linearly interpolating between the spline points. The result on a test set is shown in Fig. 12. We observe that the predicted values quite reasonably resemble the actual points. The important point to note here is that RAP discovery does not perform averaging of the data. The presence of noncentralized background noise, skewed towards the upper portion of the sample space, forces the mean to occur substantially above the actual curve.

To demonstrate this phenomenon, a three-layer feedforward neural network using a mean-square error objective function was trained with this data using the Levenberg-Marquardt algorithm. Even when a large number of hidden units is commissioned to learn the data, the sum-of-squared error could not be further reduced (see Fig. 13). In fact, with larger networks, over-fitting was observed, resulting in larger errors. This is attributed to the skewness in the data. It is, however, conceivable that a nonaveraging objective may produce more accurate results.

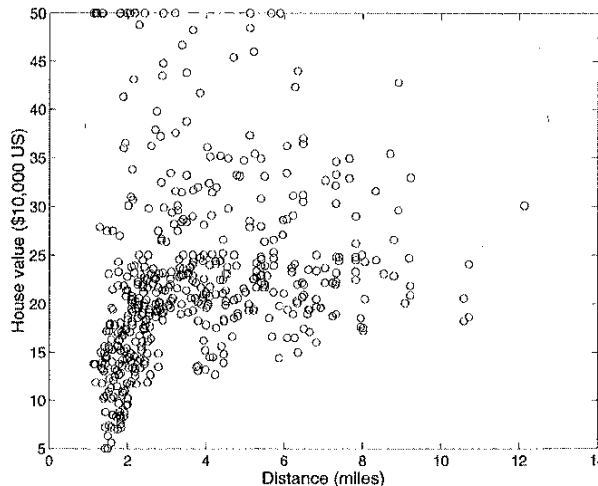


Fig. 14. Housing data.

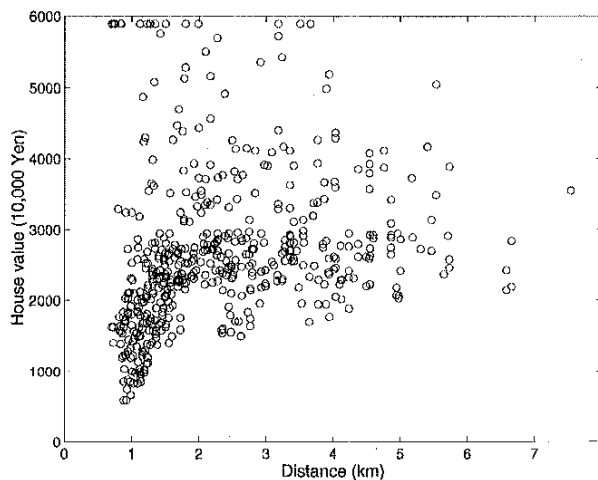


Fig. 15. Scaled housing data.

From this example, we observe the important property of robust discovery of nonlinear functional structure in the presence of noncentralized noise.

6.1.3 Scale Invariance

Next, we exemplify the important property of scale invariance. Consider the housing data shown in Fig. 14 in miles and U.S. dollars. The same data is plotted in Fig. 15, but in kilometers and Japanese yen. This latter data set will be referred to as the *scaled* housing data. Each sample point represents the value of a home located at a certain distance from the city center. When the RAP algorithm was applied, 13 events were discovered in each of the scaled and unscaled data sets. Fig. 16 and Fig. 17 show the discrete PDFs based upon the discovered events. We observe that the shape of the density function is preserved with scaling. This suggests that the relative magnitude of event density values remains constant, a necessary and sufficient condition for scale invariance. To demonstrate the usefulness of scale invariance, consider the following hypothetical query:

TABLE 2
Ratio of Probability Density Values Before and After Data Scaling $\left(\frac{Home A}{Home B}\right)$

Method	Before scaling		After scaling	
	Home A (7.08 miles, \$197,132) Home B (3.66 miles, \$417,045)	More likely	Home A (4.4 km, 23,243,834 Yen) Home B (2.27 km, 49,173,776 Yen)	More likely
RAP	2.2	A	2.2	A
PNN ₁	7.9	A	0.35	B
PNN ₂	1.05	A	1.05	A

Is it more likely to find a \$197,132 home that is 7.08 miles from the city center or a \$417,045 home that is 3.66 miles from the city center?

Ideally, we should arrive at the same answer, regardless of how the distance (kilometers or miles) and home value (dollars or yen) are specified. In fact, using the pattern discovery PDFs to answer this query, we find that the first

home is 2.2 times more likely, regardless of how distance or home value are specified.

In contrast, methods based on distance measures are generally scale sensitive. As an example, consider the probabilistic neural network (PNN) that uses the Euclidean distance. The PNN is chosen for this illustration because like RAP, it *directly* yields an axiomatically true PDF. Table 2 indicates that a seemingly harmless change of scale has reversed the decision of the PNN (PNN₁). Hopefully, the prudent practitioner would normalize the data or optimize the PNN smoothing parameter to mitigate the effects of scale. The consequence of normalizing both data sets is shown in the last row (PNN₂). The decision is now invariant to scaling and is consistent with that of RAP. The advantage of RAP is that it can avoid the often ad hoc standardization of feature variables as it is not sensitive to different magnitudes in the features and provides consistent discovery results regardless of arbitrary changes of scale.

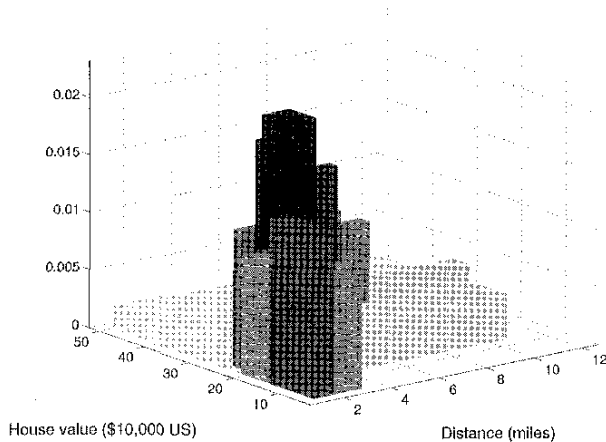


Fig. 16. Discrete PDF for housing data.

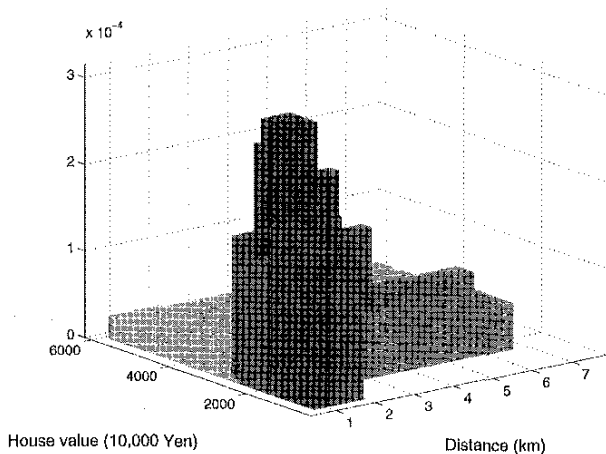


Fig. 17. Discrete PDF for scaled housing data.

6.1.4 Cluster Detection

The last example of this section is concerned with two vital properties of a discovery system, namely interpretability and discovery of subtle organization. In particular, two alternatives for interpretation are exemplified.

Consider the plasma lipid data shown in Fig. 18 and Fig. 19 taken from Scott [1]. Each point represents a measurement of 2 lipid concentrations, triglycerides, and cholesterol, taken from male subjects. It is interesting to note that the data for the diseased subjects is actually multimodal [1], although visually it appears unimodal. Further, the data for diseased and healthy subjects seems to completely overlap. The organization in this data is undoubtedly obscure and subtle.

Upon application of RAP pattern discovery, we arrive at 18 events for the diseased group and five for the healthy group, as shown in Fig. 20 and Fig. 21. To obtain a smooth representation of the discrete probability density, Gaussian kernels weighted by their normalized densities¹ are placed on the event centroids. The covariance of each kernel is computed using the points contained in each event. The result is Fig. 22. Bimodality is unambiguously revealed. Further, note that the healthy and diseased groups are now clearly delineated even though the raw data implied complete overlap.

1. The normalized density value is $\frac{\hat{p}_i}{\sum_i \hat{p}_i}$

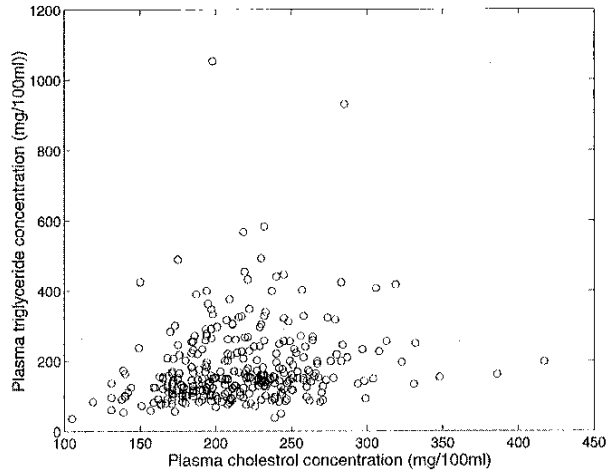


Fig. 18. Lipid data—diseased subjects.

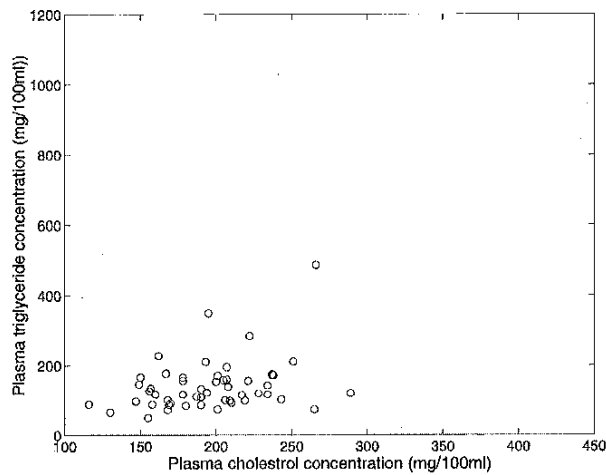


Fig. 19. Lipid data—healthy subjects.

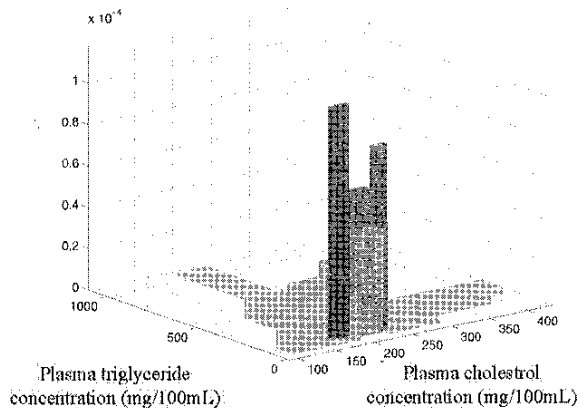


Fig. 20. Discrete PDF—diseased subjects.

In terms of interpretability, the discovered events can also be viewed as rules which characterize each group. For example, the following rules are derived from the most significant events in each group:

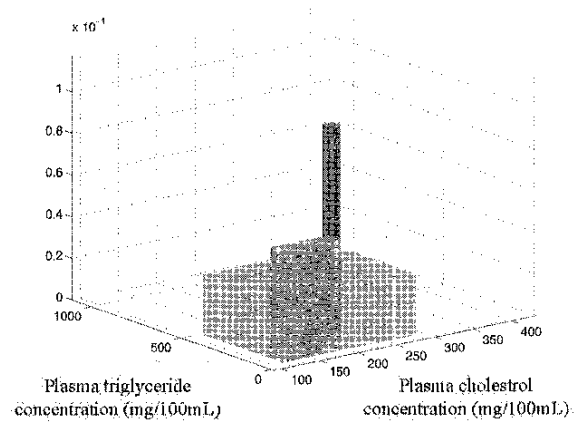


Fig. 21. Discrete PDF—healthy subjects.

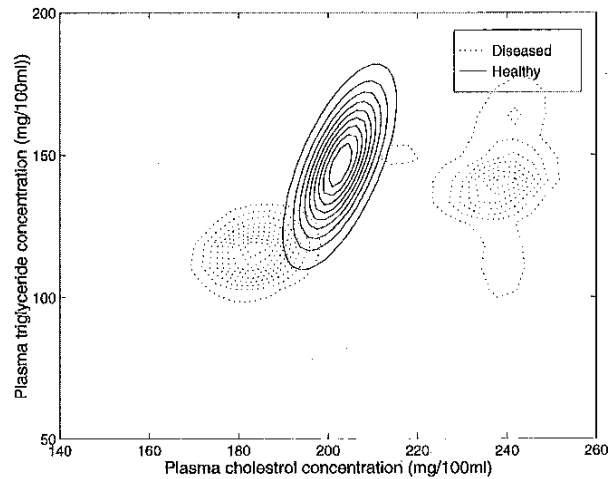


Fig. 22. Contours of smoothed events.

- If the cholesterol level is between 194 and 208 mg/100mL AND the triglyceride level is between 121 and 160 mg/100mL, then the subject is in the healthy group (supported by 10 percent of the healthy cases).
- If the cholesterol level is between 172 and 194 mg/100mL AND the triglyceride level is between 102 and 126, then the subject is in the disease group (supported by 6 percent of the disease cases).

For the disease data, cluster centers discovered by a self-organizing map (SOFM) and a mixture of Gaussian kernels [46] are shown in Fig. 23. The SOFM has been included because it is a powerful unsupervised learning algorithm and like RAP, does not require a priori information about the number of clusters. Using topographic maps ranging from 4×4 to 6×6 neurons, the SOFM finds multiple centers. The centers are determined as the weighted average of the coordinates of the neurons with the highest frequencies of activation. The weights are simply the frequency proportions of the neurons. The numerical values of the centers are provided in Table 3. With the bimodal disease data, the results of RAP pattern

TABLE 3
Clustering Results of Various Methods on Lipid Data

Method	Diseased subjects		Healthy subjects	
	# clusters detected	cluster centres	# clusters detected	cluster centres
RAP	2	(238, 141) (183, 118)	1	(203, 147)
SOFM ¹	2 (6 × 6 grid, neighborhood=1)	(223, 139) (185, 118)	1 (4 × 4 grid, neighborhood=1)	(192, 112)
Gaussian kernels	2	(233,145) (185,122)	1	(195,122)

Averaged over 10 trials.

discovery agree closely to both the findings of Scott [46], [1] and the results of the SOFM. With the healthy subjects, however, RAP is handicapped by the sparsity of the data, forbidding repeated statistical testing. This accounts for the discrepancies in those cluster centers.

RAP pattern discovery does offer some advantages. With Scott's kernel method, 370 kernels are required to furnish continuous densities for the two subject groups, while RAP only needs a total of 23. The evaluation of PDF values is therefore much quicker with the compact representation of RAP. RAP also trains faster than the SOFM, revealing cluster centers in a mere matter of seconds.

In short, we have shown that RAP can detect subtle multimodal structure without a priori knowledge of the number of clusters. The results of discovery can be easily interpreted as continuous contours or as simple rules and are shown to be consistent with those of existing methods.

6.2 Multidimensional Data

In the previous section, we illustrated a number of basic properties of RAP pattern discovery. Now, we demonstrate the usefulness of some of these properties in a real-life, multidimensional data set. We consider the fairly

well-understood area of thyroid gland disease to allow for easy verification of the discovered patterns. The data is taken from Coomans et al. [47] and consists of three categories, hypothyroidism (deficiency in thyroid hormones), hyperthyroidism (excess of thyroid hormones) and euthyroidism (normal thyroid function). The features are five continuous clinical measurements summarized in Table 4. The data for each class is displayed using a parallel axes plot [48] in Fig. 24 and Fig. 25. The application of RAP pattern discovery yielded five significant events for euthyroidism, two for hyperthyroidism, and three for hypothyroidism subjects. To display multidimensional events we use a slightly modified parallel axes plot. As usual, the d -axes are drawn parallel and equally spaced. A d -dimensional event is represented by connecting the endpoints of each one-dimensional interval by a piecewise linear curve. The most significant event in each of the three categories is shown in Fig. 26, and the upper portion is magnified in Fig. 27. These plots lend to rapid verification with domain theory. Consider the following observations:

- The features T4 and T3RIA represent the two types of hormones produced by the thyroid gland. As expected, the events suggest that the hypothyroid subjects exhibit a reduced level of these hormones while the hyperthyroid subjects are characterized by abnormally high hormone concentrations. The euthyroid hormone levels fall between those of the two disease states.
- The events also suggest that TSH is substantially higher in the hypothyroid case. Again, this is in agreement with physiological principles. In hypothyroidism, the low-functioning thyroid gland produces very few thyroid hormones (T3 and T4). The negative feedback that these hormones normally provide to regulate TSH production is removed and TSH levels soar. The opposite situation occurs in the hyperthyroid case and is also elucidated by the significant events.

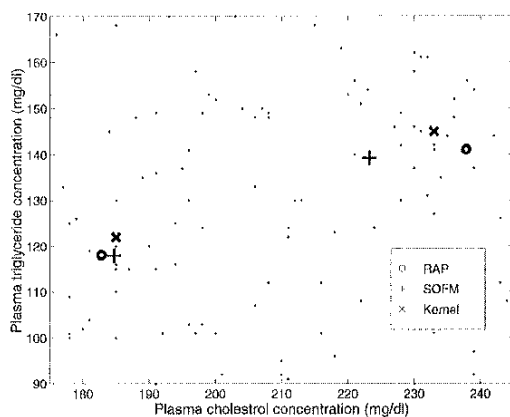


Fig. 23. Discovered cluster centers.

TABLE 4
Thyroid Data Features

Feature number	Measurement	Units	Symbol
1	T3-resin uptake test	%	RT3U
2	Total serum thyroxin	$\mu\text{g/dL}$	T4
3	Total serum triiodothyronine	$\mu\text{g/L}$	T3RIA
4	Basal thyroid-stimulating hormone	$\mu\text{IU/mL}$	TSH
5	Maximal absolute difference of the TSH value after injection of 200 μg of thyrotropin-releasing hormone (TRH) as compared to the basal value.	$\mu\text{IU/mL}$	dTSH

- The observation that

dTSH (hypothyroid)
> dTSH (euthyroid)
> dTSH (hyperthyroid)

can be explained with similar reasoning. Note that TRH is a hormone which stimulates the production of TSH. In the hyperthyroid case, there is so much inhibition from the excess thyroid hormones, that additional TRH has little effect on TSH production. The opposite is true for hypothyroidism. In the

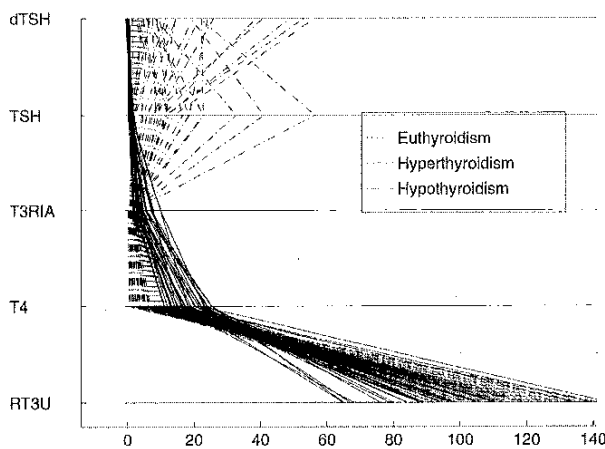


Fig. 24. Parallel axes plot of thyroid data.

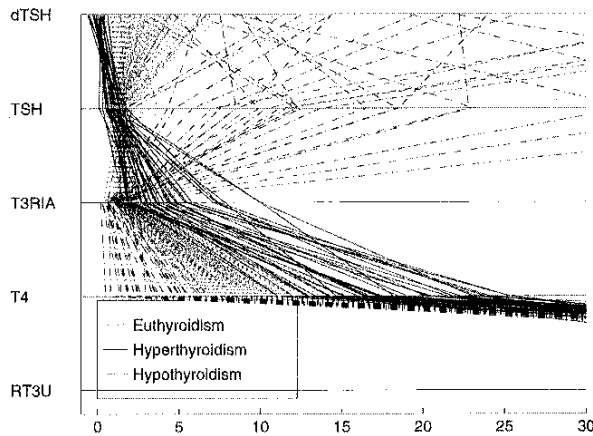


Fig. 25. Zoom-in on features 2-5.

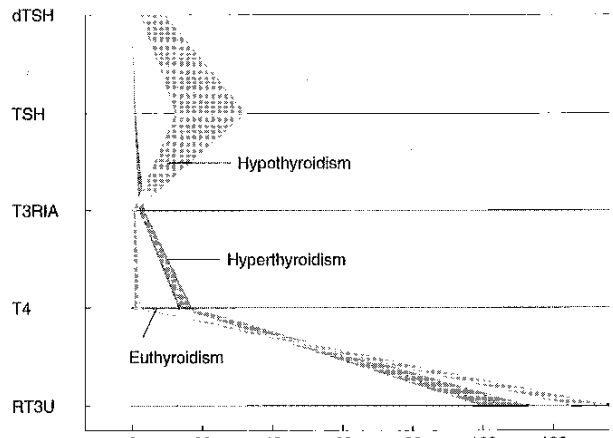


Fig. 26. Most significant events.

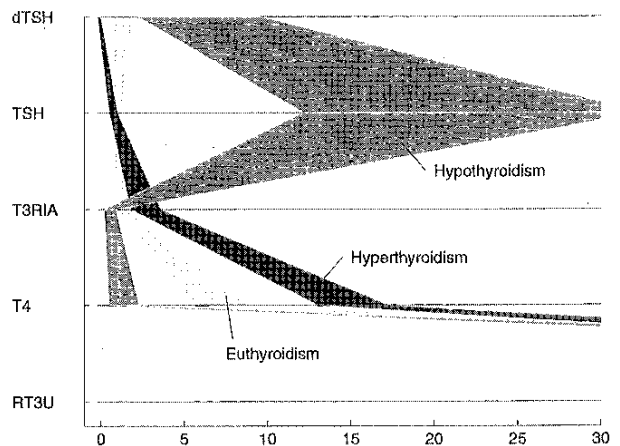


Fig. 27. Magnified significant events.

TABLE 5
Event Statistics for the Most Significant Events

Category	Event #	Residual	% Support
Euthyroidism	1	2.7	5.3
Hyperthyroidism	1	2.3	14.3
Hypothyroidism	2	2.6	20.0

normal functioning thyroid, TRH does end up increasing the production of T3 and T4, but their negative feedback counteracts the effect of TRH.

Conventional exploratory data analysis, such as parallel axes plots, may qualitatively hint at the separable structure in the T4, T3RIA, and TSH features. However, apart from quantifying this information, pattern discovery also sheds light on the previously obscured structure of the dTSH and RT3U features. The plot of events has two advantages over conventional parallel axes plots of the data. The problem of splotching [1] is overcome by only showing significant events. Unlike thinning procedures [1], here there is no fear of losing valuable structural information. Secondly, the event plot implicitly carries useful quantitative information about the discovered structure, in terms of event boundaries, significance, and percent support of each event (Table 5). It is worthwhile to mention that the most significant event need not necessarily be the most frequently supported event. Indeed, significance implies the existence of organization which is not conveyed by simple frequency statistics. As a last remark, we note that, as in the housing example, here, the discovered events can also be easily interpreted as logical rules.

In this last experiment, we have demonstrated the application of RAP pattern discovery to multidimensional data. Relationships in the data are made apparent by uncovering significant events. The parallel axes event plot is uncluttered and facilitates interpretation. The discovered events can be easily validated with available domain theory and provide quantitative information about the data's organization.

7 CONCLUDING REMARKS

7.1 Limitations

For k levels of recursion, partitioning with a fixed Q will yield $(Q^d)^k$ candidate events for examination. Clearly, for $d > 5$, the number of computations becomes prohibitive even for small Q . In such instances, the partitioning approximation must be abandoned in favour of a mathematical optimization approach to find significant events in a compact region of the d dimensional space. We stress that the theoretical formulation of significant events via contingency tables applies regardless of dimensionality. In fact, for any value of $d > 2$, the analysis is always two-dimensional. Only the partitioning approximation to the

optimization problem needs to be replaced by an unconstrained optimization algorithm.

A second limitation of partitioning is the inability to directly handle discrete data that have only a few discrete levels. Data preprocessing would be required but such operations may detract from the interpretability of pattern discovery.

7.2 Summary

This paper has presented a novel approach to pattern discovery based on the theoretical formulation of a contingency table of events. A framework of events was put forth and discovery was posed as residual analysis. The distinguishing characteristic of RAP pattern discovery is its wide range of applicability and straightforward interpretability. The same algorithm that is used for uncovering functional structure is readily applied to the detection of multiple modes and interpretation of multi-dimensional data.

With real-life data, it is often difficult to ascertain the suitability of a continuous function model or the number of clusters or the type of embedded noise. Often no a priori evidence is available to even suggest that the underlying structure is spatially contiguous. Unlike the majority of model-based (statistical) and model-free (neural) discovery methods, RAP pattern discovery thrives when these assumptions cannot be readily made. The experiments demonstrated several strengths of the pattern discovery method, including tolerance of non-centralized noise, scale invariance, and rapid training. One of the most significant advantages of RAP over model-free discovery methods is that the results are readily interpreted, either by way of simple rules, contour plots, or parallel event plots.

APPENDIX A

1. TERMINATION CONDITION

In the derivation of the large sample variance of the residual [29], [27], [30], [24], the central limit theorem is applied to the summation of multinomials, yielding an asymptotically normal distribution. For this approximation to hold, the expected cell frequency should be sufficiently large, typically at least 25 samples. Let ζ_{min} denote this minimum expected cell frequency. We therefore constrain the expected frequency given by (8),

$$\hat{m}_j = \frac{1}{2}(n_{1j} + n_{2j}) \geq \zeta_{min} \quad (25)$$

If we substitute for n_{2j} using (3) and solve for the event volume, v_j , we arrive at a lower bound for v_j .

$$v_j \geq \frac{V_{TOT}}{n_{1+}}(2\zeta_{min} - n_{1j}) = \text{lower bound} \quad (26)$$

If the volume of an event falls below this lower bound, partitioning should not proceed. This condition is expressed by the second termination criterion.

2. SCALE INVARIANCE

We verify (22) and (23). For the sake of clarity, variables will be written explicitly in terms of events. By the frequency equalization tendency of MMEP, we see that along the k th dimension of Ω' , the partition points will occur at

$$a'_1 = \gamma_1 a_1, a'_2 = \gamma_2 a_2, \dots, a'_k = \gamma_k a_k \quad (27)$$

where a_1, a_2, \dots, a_d are the partition points along the k th dimension of Ω . Extending this argument to every dimension, $k = 1, \dots, d$, we immediately see that event volumes are related as,

$$v(L'_j) = \Gamma v(E_j) \quad (28)$$

where

$$\Gamma = \prod_{k=1}^d \gamma_k.$$

The frequency equalization tendency also ensures that event frequencies are invariant,

$$n(E'_j) = n(E_j) \quad (29)$$

This result directly verifies (22). To verify (23), simply substitute the density definition (18) into both sides of (23). Simplifying, we arrive at,

$$\frac{n(E_i) v(E_j)}{v(E_i) n(E_j)} = \frac{n(E'_i) v(E'_j)}{v(E'_i) n(E'_j)} \quad (30)$$

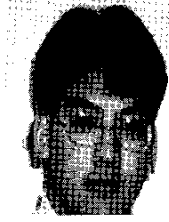
Using the relations (28) and (29), the result (23) follows immediately.

REFERENCES

[1] D.W. Scott, *Multivariate Density Estimation*. John Wiley and Sons, 1992.
 [2] P.J. Huber, "Projection Pursuit (with Discussion)," *Annals of Statistics*, vol. 13, pp. 435-525, 1985.
 [3] M. Rosenblatt, "Remarks on Some Nonparametric Estimates of a Density Function," *Annals of Math. Statistics*, vol. 27, pp. 832-837, 1956.
 [4] E. Parzen, "On Estimation of Probability Density Function and Mode," *Annals of Math. Statistics*, vol. 33 pp. 1,065-1,076, 1962.
 [5] J. Moody and C.J. Darken, "Fast Learning in Networks of Locally-Tuned Processing Units," *Neural Computation*, vol. 1, no. 2, pp. 281-294, 1989.
 [6] R.O. Duda and P.H. Hart, *Pattern Classification and Scene Analysis*. John Wiley and Sons, 1973.
 [7] E.J. Hartman, J.D. Keeler, and J.M. Kowalski, "Layered Neural Networks with Gaussian Hidden Units as Universal Approximations," *Neural Computation*, vol. 2, no. 2, pp. 210-215, 1990.
 [8] J. Park and I.W. Sandberg, "Approximation and Radial Basis Function Networks," *Neural Computation*, vol. 5, no. 2, pp. 305-316, 1993.
 [9] C. Bishop, *Neural Networks for Pattern Recognition*, Oxford Univ. Press, 1995.
 [10] D.E. Rumelhart, G.E. Hinton, and R.J. Williams, "Learning Representations by Backpropagating Errors," *Nature*, vol. 323, pp. 533-536, 1986.
 [11] G. Cybenko, "Approximation by Superpositions of a Sigmoidal Function," *Math. of Control, Signals, and Systems*, vol. 2, pp. 303-314, 1989.
 [12] H. White, "Learning in Artificial Neural Networks: A Statistical Perspective," *Neural Computation*, vol. 1, pp. 425-464, 1989.

[13] G.E. Hinton, J.L. McClelland, and D.E. Rumelhart, "Distributed Representations," *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, D.E. Rumelhart and J.L. McClelland, eds., vol. 1, MIT Press, Cambridge, Mass., 1986.
 [14] J. Wejchert and G. Tesauro, "Visualizing Processes in Neural Networks," *IBM J. Research and Development*, vol. 35, pp. 244-253, 1991.
 [15] H. Narazaki, M. Yamamoto, and T. Watanabe, "Reorganizing Knowledge in Neural Networks: An Explanation Mechanism for Neural Networks in Data Classification Problems," *IEEE Trans. Systems, Man, and Cybernetics*, part B, vol. 26, no. 1, pp. 107-117, 1996.
 [16] S. Avner, "Extraction of Comprehensive Symbolic Rules from a Multilayer Perceptron," *Eng. Applications*, vol. 9, no. 2, pp. 137-43, 1996.
 [17] R. Andrews, R. Cable, J. Diederich, S. Geva, M. Golea, R. Hayward, C. Ho-Stuart, and A. Tickle, "Survey and Critique of Techniques for Extracting Rules from Trained Artificial Neural Networks," *Knowledge-Based Systems*, vol. 8, no. 6, pp. 373-389, 1995.
 [18] M. Dolson, "Discriminative Nonlinear Dimensionality Reduction for Improved Classification," *Int'l J. Neural Systems*, vol. 5, no. 4, pp. 313-333, 1994.
 [19] T. Kohonen, *Self-Organizing Maps*. Springer, Berlin, 1995.
 [20] B.D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge Univ. Press, 1996.
 [21] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. McGraw-Hill, 1991.
 [22] A.K.C. Wong and Y. Wang, "High-Order Pattern Discovery from Discrete-Valued Data Sets," *IEEE Trans. Knowledge and Data Eng.*, pp. 877-893, vol. 9, no. 6, Nov./Dec. 1997.
 [23] S.C. Port, *Theoretical Probability for Applications*. John Wiley and Sons, 1994.
 [24] R.A. Christensen, *Log-Linear Models*. Springer-Verlag, 1990.
 [25] S. Haberman, *The Analysis of Qualitative Data*, vol. 1. Academic Press, 1978.
 [26] B.S. Everitt, *The Analysis of Contingency Tables*. Wiley, New York, 1977.
 [27] A. Agresti, *Categorical Data Analysis*. Wiley, New York, 1990.
 [28] R.N. Forthofer, *Public Program Analysis: A New Categorical Data Approach*. Lifetime Learning Publications, 1981.
 [29] S.J. Haberman, *The Analysis of Frequency Data*. Univ. of Chicago Press, 1974.
 [30] C. Cox, "An Elementary Introduction to Maximum Likelihood Estimation for Multinomial Models: Birch's Theorem and the Delta Method," *Amer. Statistician*, vol. 38, no. 4, pp. 283-287, 1984.
 [31] A.K.C. Wong, D.K.Y. Chiu, and B. Cheung, "Information Discovery Through Hierarchical Maximum Entropy Discretization," *Knowledge Discovery in Databases*, G. Piatetsky-Shapiro and W.J. Frawley, eds., pp. 125-140, AAAI/MIT Press, 1987.
 [32] M. Lascurain, "On Maximum Entropy Discretization and its Applications in Pattern Recognition," PhD thesis, Systems Design Engineering, Univ. of Waterloo, Waterloo, Ont., Canada, 1983.
 [33] C.E. Shannon, "Mathematical Theory of Communication," *Bell Systems Technical J.*, vol. 27, no. 3, pp. 379-423, 1948.
 [34] E.T. Jaynes, "Information Theory and Statistical Mechanics," *Physical Rev.*, vol. 106, no. 1, pp. 620-630, 1957.
 [35] D.K.Y. Chiu, B. Cheung, and A.K.C. Wong, "Information Synthesis Based on Hierarchical Maximum Entropy Discretization," *J. Experimental and Theoretical Artificial Intelligence*, vol. 2, pp. 117-129, 1990.
 [36] J. Sonquist, *Multivariate Model Building: The Validation of a Search Strategy*, Inst. for Social Research, Univ. of Michigan, Ann Arbor, 1970.
 [37] L. Gordon and R.A. Olshen, "Asymptotically Efficient Solutions to the Classification Problem," *Annals of Statistics*, vol. 6, no. 3, pp. 515-533, 1978.
 [38] J.H. Friedman, "A Recursive Partitioning Decision Rule for Nonparametric Classification," *IEEE Trans. Computers*, pp. 404-408, 1977.
 [39] E.G. Henrichon and K.S. Fu, "A Nonparametric Partitioning Procedure for Pattern Classification," *IEEE Trans. Computers*, vol. 18, no. 7, pp. 614-624, 1969.
 [40] A. Ciampi, C.H. Chang, S.A. Hogg, and S. McKinney, "Recursive Partition: A Versatile Method for Exploratory Data Analysis in Biostatistics," *Biostatistics*, I.B. MacNeil and G.J. Umphrey, eds., vol. 5, pp. 23-50, Dordrecht, Netherlands, 1987.

- [41] D.K.Y. Chiu and A.K.C. Wong, "Synthesizing Knowledge: A Cluster Analysis Approach Using Event Covering," *IEEE Trans. Systems, Man, and Cybernetics*, vol. 16, no. 2, pp. 251-259, 1986.
- [42] R. Fletcher, *Practical Methods of Optimization*. John Wiley and Sons, second ed., 1987.
- [43] T.W. Anderson, "Some Nonparametric Multivariate Procedures Based on Statistically Equivalent Blocks," *Multivariate Analysis*, P.R. Krishnaiah, ed., pp. 5-28, Academic Press, 1966.
- [44] L.P. Devroye, "A Universal k-Nearest Neighbor Procedure in Discrimination," *Proc. IEEE Computer Soc. Conf. Pattern Recognition and Image Processing*, pp. 142-147, 1978.
- [45] W.S. Meisel and D.A. Michalopoulos, "A Partitioning Algorithm with Application in Pattern Classification and the Optimization of Decision Trees," *IEEE Trans. Computers*, vol. 22, no. 1, pp. 93-103, 1973.
- [46] D.W. Scott, A.M. Gotto, J.S. Cole, and G.A. Gorry, "Plasma Lipids as Collateral Risk Factors in Coronary Artery Disease—A Study of 371 Males with Chest Pain," *J. Chronic Diseases*, vol. 31, pp. 337-345, 1978.
- [47] D. Coomans, I. Broeckaert, M. Jonckheer, and D.L. Massart, "Comparison of Multivariate Discrimination Techniques for Clinical Data," *Methods of Information in Medicine*, vol. 22, pp. 93-101, 1983.
- [48] E.J. Wegman, "Hyperdimensional Data Analysis Using Parallel Coordinates," *J. Amer. Statistical Assoc.*, vol. 85, no. 411, pp. 664-675, 1990.



Tom Chau (S'92-M'97) received the BAsC degree in engineering science in 1992 and the MASc degree in electrical engineering in 1994, both from the University of Toronto. In 1997, he received the PhD degree in systems design engineering at the University of Waterloo, Canada. Since 1996, he has been a consultant with KPMG and, subsequently, with IBM Canada. He recently joined the research staff at the Bloorview MacMillan Centre, a world-leading pediatric rehabilitation research institution in Toronto. He is also an adjunct assistant professor at the Institute of Biomaterials and Biomedical Engineering at the University of Toronto. His research interests include statistical theory of pattern discovery, rehabilitation engineering, gesture recognition, and quantitative gait analysis. He is a member of the IEEE and the IEEE Engineering in Medicine and Biology Society.



Andrew K.C. Wong (M'79) received his PhD degree from Carnegie Mellon University, Pittsburgh, Pennsylvania, in 1968 and, afterward, taught there for several years. He is currently a full professor of systems design engineering and director of the Pattern Analysis and Machine Intelligence (PAMI) Laboratory at the University of Waterloo, Canada. He is also an honorable professor at the University of Hull, United Kingdom. He has authored and coauthored chapters and sections in a number of books on engineering and computer science, and has published many articles in scientific journals and conference proceedings. He is the 1991 recipient of the Federation of Chinese Canadian Professionals Award of Merit, and he is a member of the IEEE.