Pattern Recognition Letters

Authorship Confirmation

Please save a copy of this file, complete and upload as the "Confirmation of Authorship" file.

As corresponding author I, Robert Haralick, hereby confirm on behalf of all authors that:

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.
2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.
3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.
4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.
5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

SignatureRobert M. Haralick DateApril 3, 2017

Graphical Abstract (Optional)

To create your abstract, please type over the instructions in the template box below. Fonts or abstract dimensions should not be changed or altered.

Dependence

Robert M. Haralick

This paper discusses different kinds of dependency. For numerically valued variables our discussion centers on the maximal correlation coefficient and its cousin the monotone correlation coefficient. We show how to calculate the maximal correlation coefficient in the case the random variables take on a finite set of values. For non-numerically valued variables our discussion centers on information theoretic measures related to mutual information and we describe some that are also metrics. We visually illustrate the difference between these two kinds of measures with a texture example that computes the joint probability image: an image in which the gray level of each pixel is the joint probability of the gray levels of the pixels in its neighborhood. Neighborhoods can be regular such as 5×5 or they can be irregular. Finally, we discuss manifold methods for classification: the N-tuple method, the subspace classifiers, the subspace ensemble classifiers, including the graphical model for representing the class conditional probability distribution. We describe a procedure to convert an N-tuple classifier to a graphical model classifier. We also conjecture that there is new form of a universal approximation theorem by which not too complex classification functions from measurement space to the set of classes can be approximately represented in the form of a subspace classifier using multiple subspaces such as the N-tuple method.

Research Highlights (Required)

To create your highlights, please type the highlights against each `\item` command.

It should be short collection of bullet points that convey the core findings of the article. It should include 3 to 5 bullet points (maximum 85 characters, including spaces, per bullet point.)

- Maximal Correlation Coefficient and Monotone Correlation Coefficient
- Information Theoretic: Mutual Information, its cousins and metric forms
- Cooccurrence Probabilities and the Joint Probability Image
- Manifold Methods: the N-tuple method, the subspace methods, and subspace ensemble methods

Pattern Recognition Letters
journal homepage: www.elsevier.com

Dependence

Robert M. Haralick[a,**]

[a]*Graduate Center, City University of New York, 365 Fifth Avenue, New York, NY, 10116, United States*

ABSTRACT

This paper discusses different kinds of dependency. For numerically valued variables our discussion centers on the maximal correlation coefficient and its cousin the monotone correlation coefficient. We show how to calculate the maximal correlation coefficient in the case the random variables take on a finite set of values. For non-numerically valued variables our discussion centers on information theoretic measures related to mutual information and we describe some that are also metrics. We visually illustrate the difference between these two kinds of measures with a texture example that computes the joint probability image: an image in which the gray level of each pixel is the joint probability of the gray levels of the pixels in its neighborhood. Neighborhoods can be regular such as 5×5 or they can be irregular. Finally, we discuss manifold methods for classification: the N-tuple method, the subspace classifiers, the subspace ensemble classifiers, including the graphical model for representing the class conditional probability distribution. We describe a procedure to convert an N-tuple classifier to a graphical model classifier. We also conjecture that there is new form of a universal approximation theorem by which not too complex classification functions from measurement space to the set of classes can be approximately represented in the form of a subspace classifier using multiple subspaces such as the N-tuple method.

1. Introduction

There is much research using measures of dependency to discover associations between variables. When there are large numbers of variables, the natural methodology is to evaluate a measure of dependence between each pair of variables, sort the pairs of variables from highest dependence to lowest dependence and then try to understand why certain pairs of variables have high dependence and others low dependence. The deepest level of such understanding would be to use the joint relationship among the variables to construct a model that predicts what had been observed.[1]

There are many measures of statistical dependency between variables. It would take a long paper to survey them all. Our purpose is to discuss a few kinds of dependence. Before we start we make mention without elaboration of the classic paper on dependency in contingency tables: the article by (Goodman and Kruskal, 1954).

There are many forms of dependency in Pattern Recognition. There is the basic dependency between pairs of variables. There is the dependency between subsets of variables that enable dimensionality reduction and manifold learning. There is the dependency between the independent variables and the response variable in prediction tasks. There is the dependency between the measurement tuples and true class in the the classifier task.

We will begin with the basic dependency between variables, discussing what we regard as the best association measures for numerically valued variables and then methods that can take care of numerically valued or categorically valued variables. We will illustrate how in image data, cooccurrence probabilities can be used to measure the strength of dependence in image neighborhoods and thereby distinguish the difference between low joint probability neighborhoods and high joint probability neighborhoods. Remarkably, low joint probability neighborhoods will correspond with edge regions. Finally, we explore some aspects of dependency through subspaces. We explore the N-tuple method, subspace classifiers, and subspace ensemble

[**]Corresponding author: Robert Haralick Tel.: +1-212-817-8192; Fax: +1-212-817-1510;

e-mail: rharalick@gc.cuny.edu (Robert M. Haralick)

[1] However, with observational data, there are certainly multiple models that can equally well explain the data.

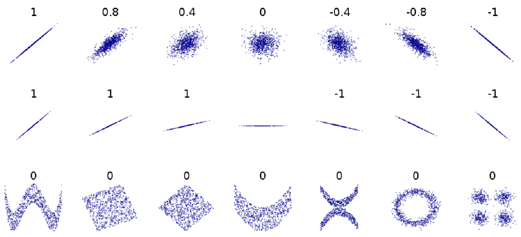


Fig. 1: Graphics from Wikipedia article on correlation and dependence.

classifiers. We conjecture that there is new form of a universal approximation theorem by which not too complex classification functions, defined from measurement space to the set of classes, can be approximately represented in the form of a subspace classifier using multiple subspaces such as the N-tuple classifier and the subspace ensemble classifiers. If this is so, it would imply that the way to think about classifiers is through subspaces and sub-manifolds.

There are two kinds of tasks in characterizing dependency: measuring the strength of the dependency between/among variables and estimating the constraint that defines the dependency.

The most common measure of dependency is the correlation coefficient. For numerically valued random variables X and Y, the correlation coefficient is defined by

$$\rho(X,Y) = E_{XY}\left[\frac{(X-\mu_x)}{\sigma_x}\frac{(Y-\mu_y)}{\sigma_y}\right]$$

Note that if $E_X[X] = E_Y[Y] = 0$ and $V_X[X] = V_Y[Y] = 1$ then $\rho(X,Y) = E_{XY}[XY]$.[2]

Figure (1) shows scattergrams of various kinds of dependencies and the value of their correlation coefficient. Notice that the correlation coefficient only measures the linear portion of the dependency.

Dependency can, of course, be more complex than linear dependency. Dependency can have multiple parts. Examine the dependency in Figure (2). This kind of multi-part dependency is not captured by any of the dependency measures and must be first processed by a non-linear manifold clustering algorithm.

2. Rènyi's Conditions

One of the earliest people to specify a list of properties that a measure of dependency should have was (Rényi, 1959). Rènyi wanted to characterize by a numerical value, the strength of the dependence between two random variables. He wanted the measure to be normalized in the interval [0, 1], with 0 being the value associated with no dependence, i.e. independence, and 1 being the value associated with complete or strict dependence.

He made a list of properties that he thought was essential for any measure of dependency δ.

(A) $\delta(X,Y)$ is defined for any pair of random variables X, Y neither of them being constant with probability 1

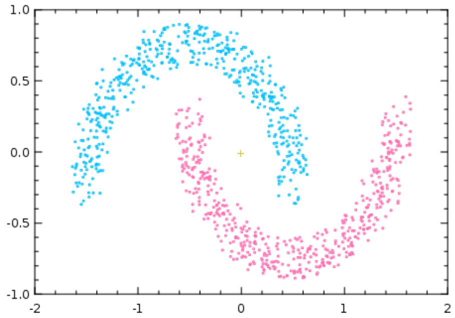


Fig. 2: For some values of X, Y has multiple values. For some values of Y, X has multiple values. There are two non-linear disconnected manifolds.

(B) $\delta(X,Y) = \delta(Y,X)$

(C) $0 \leq \delta(X,Y) \leq 1$

(D) $\delta(X,Y) = 0$ *if and only if* X and Y are independent

(E) $\delta(X,Y) = 1$ *if* Y is completely dependent on X or X is completely dependent on Y, This means that $X = g(Y)$ or $Y = f(X)$

(F) If f and g are one-to-one mappings, then $\delta(f(X), g(X)) = \delta(X,Y)$

(G) If X and Y are jointly normal, then $\delta(X,Y) = |\rho(X,Y)|$, where ρ is the correlation coefficient

Rènyi thought about changing (E) to $\delta(X,Y) = 1$ *only if* Y is completely dependent on X or X is completely dependent on Y, but he thought this to be too restrictive.

The correlation coefficient only satisfies (B), (C) in absolute value, and (G). Rènyi explored a few measures of dependency: the maximal correlation coefficient introduced by (Hirschfeld, 1935) and then by (Gebelein, 1941), the correlation ratio, and a mutual information measure.

The correlation ratio $\eta^2_{Y|X}$ of Y on X is defined by

$$\eta^2_{Y|X} = \frac{V_X[E_Y[Y|X]]}{V_Y[Y]}$$

It does not satisfy (A), (B), (D), and (F). And even if it is made symmetric

$$\eta(X,Y)^2 = max\{\eta^2_{Y|X}, \eta^2_{X|Y}\}$$

$\eta(X,Y)^2$ still does not satisfy (A), (D), and (F).

There are relationships between the correlation coefficient and the correlation ratio. Here we mention the most important two relations.

$$\rho(X,Y)^2 \leq \eta^2_{Y|X}$$

Let $G = \{g : \to\mathbb{R} \mid E[g(X)] = 0 \text{ and } V[g(X)] = 1\}$ then, for any $g \in G$,

$$\rho^2(g(X),Y) \leq \eta^2_{Y|X}$$

with equality when $g(x) = E_Y[Y|X = x]$.

Another fundamental property of the correlation ratio is

[2] We always use E to mean expectation and V to mean variance, and σ to mean $\sqrt{V}$.

$$\inf_g E_{XY}[(Y - g(X))^2] \quad = \quad 1 - \eta^2_{Y|X} \tag{1}$$

Here again, the function g that achieves the inf is

$$g(x) = E_Y[Y|X = x]$$

(Sampson, 1984) extended the idea of correlation ratio of Y on X to the multivariate case.

Definition 1. *The* **Multivariate Correlation Ratio** η_Λ *measuring the predictability of the random vector Y from the random vector X relative to the positive definite square matrix* Λ *is defined by*

$$\eta^2_\Lambda(Y|X) \quad = \quad \frac{tr\left(\Lambda^{-1}Cov_{XX}[E_Y[Y|X]]\right)}{tr\left(\Lambda^{-1}\Sigma_Y\right)}$$

Sampson uses the matrix Λ as a weight matrix. when $\Lambda = I$, the identity matrix, then it corresponds to an unweighted least squares. When Λ is diagonal, the diagonal elements are the weights for the corresponding components of Y. When $\Lambda = \Sigma_Y$ the weighting is like that used in the Mahalanobis distance. Sampson notes that his multivariate definition satisfies the multivariate equivalent of (1). Let $\| s \|^2_\Lambda = s'\Lambda^{-1}s$. Then

$$\inf_g \| E[Y - g(X) \|^2_\Lambda = \left(1 - \eta^2_\Lambda(Y;X)\right) tr\left(\Lambda^{-1}\Sigma_Y\right)$$

and this minimum occurs when $g(x) = E[Y|X = x]$.

When $\Lambda = I$, the expression for the multivariate correlation ratio simplifies.

$$\eta^2_I(Y|X) \quad = \quad \frac{tr\left(Cov_{XX}[E_Y[Y|X]]\right)}{tr\left(\Sigma_Y\right)} \tag{2}$$

(Kabe and Gupta, 1990) discuss a formulation that they claim is better than that of Sampson's. They replace the trace with the determinant. Corresponding to (2), their expression for the multivariate correlation ratio is given by (3)

$$\eta^2(Y|X) = \frac{\left|Cov_{XX}[E_Y[Y|X]]\right|}{|\Sigma_Y|} \tag{3}$$

3. Maximal Correlation Coefficient

Definition 2. *Let X and Y be numerically valued random variables. Define sets of Borel measurable functions F and G by*

$$\begin{aligned} F &= \{f : \mathbb{R} \rightarrow \mathbb{R} \mid E[f(X)] = 0; V[f(X)] = 1\} \\ G &= \{g : \mathbb{R} \rightarrow \mathbb{R} \mid E[g(Y)] = 0; V[g(Y)] = 1\} \end{aligned}$$

The **Maximal Correlation Coefficient** ρ_{max} *between X and Y is defined by*

$$\rho_{max}(X,Y) \quad = \quad \sup_{f \in F, g \in G} E_{XY}[f(X)g(Y)]$$

(Yu, 2008) notes that

$$\rho_{max}(X,Y) = \sup_{g \in G}(V_Y[E_X[g(X)|Y]])^{\frac{1}{2}}$$

And if non-degenerate random variables X and Z are conditionally independent given Y, then

$$\rho_{max}X, Z \leq \rho_{max}(X,Y)\rho_{max}(Y,Z)$$

If X and Y are bivariate normal, then $\rho_{max}(X,Y) = |\rho(X,Y)|$.

If $\rho_{max}(X,Y) = |\rho(X,Y)|$, then for some constants a_0, a_1, b_0, b_1,

$$\begin{aligned} E[X|Y] &= a_1 Y + a_0 \\ E[Y|X] &= b_1 X + b_0 \end{aligned}$$

(Lancaster, 1957) proved the following property: that if X and Y are distributed as a bivariate Gaussian, then any functional transformation of them will yield an absolute correlation no larger than the absolute correlation of X and Y. Let X and Y be distributed as a bivariate normal distribution with correlation $\rho(X,Y)$. and let

$$\begin{aligned} F &= \{f : \mathbb{R} \rightarrow \mathbb{R} \mid E[f(X)] = 0; V[f(X)] = 1\} \\ G &= \{g : \mathbb{R} \rightarrow \mathbb{R} \mid E[g(Y)] = 0; V[g(Y)] = 1\} \end{aligned}$$

then for any $f \in F$ and $g \in G$

$$|\rho(f(X), g(Y))| \leq |\rho(X,Y)|$$

There is a relationship between the maximal correlation coefficient and the Cramér's V statistic which is based on the χ^2 test for independence in contingency tables. The Cramér's V statistic is a measure of dependency in a $J \times K$ contingency table.

$$\chi^2 \quad = \quad \sum_{j=1}^{J}\sum_{k=1}^{K} \frac{\left(n_{jk} - \frac{n_{j.}n_{.k}}{n_{..}}\right)^2}{\frac{n_{j.}n_{.k}}{n_{..}}} \tag{4}$$

$$\text{Cramer's } V \quad = \quad \sqrt{\frac{\chi^2/n_{..}}{min\{J-1, K-1\}}} \tag{5}$$

(Gautam and Kimeldorf, 1999) develop this relationship for the $2 \times K$ contingency table as shown in Figure (3). For $J = 2$, $V = \sqrt{\chi^2/n_{..}}$. Gautam and Kimeldorf showed that in this case

$$\rho_{max}(Y,X) \quad = \quad \sup_f \rho(Y, f(X)) = V$$

The function f that achieves the *sup* is $f(k) = n_{.k}$.

The maximal correlation coefficient is difficult to determine in the continuous case. But for the finite case there is a method mentioned by (Haralick et al., 1973) and also by (Witsenhausen, 1975) using an eigenvector eigenvalue method. Here we show a singular value decomposition method.

Let X take values from the ordered set $\{\alpha_1, \ldots, \alpha_I\}$. Let Y take values from the ordered set $\{\beta_1, \ldots, \beta_J\}$. Define $p_{ij} =$

Categories							
	X = 1	**X = 2**	...	**X = k**	...	**X = K**	**Total**
Score	$f(1)$	$f(2)$	...	$f(k)$	...	$f(K)$	
Y = 1	n_{11}	n_{12}	...	n_{1k}	...	n_{1K}	$n_{1.}$
Y = 2	n_{21}	n_{22}	...	n_{2k}	...	n_{2K}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	...	$n_{.K}$	$n_{..}$

Fig. 3: The General $2 \times K$ Contingency Table

$P(X = \alpha_i, Y = \beta_j)$ and define

$$\begin{aligned} p_{i.} &= \sum_{j=1}^{J} p_{ij} \\ p_{.j} &= \sum_{i=1}^{I} p_{ij} \end{aligned}$$

Find new values $a_i = f(\alpha_i)$ and $b_j = g(\beta_j)$ satisfying

- $\mu_{fx} = E[f(X)] = \sum_{i=1}^{I} a_i p_{i.} = 0$
- $\sigma^2_{fx} = E[(f(X) - \mu_{fx})^2] = \sum_{i=1}^{I} a_i^2 p_{i.} = 1$
- $\mu_{gy} = E[g(Y)] = \sum_{j=1}^{J} b_j p_{.j} = 0$
- $\sigma^2_{gy} = E[(g(Y) - \mu_{gy})^2] = \sum_{j=1}^{J} b_j^2 p_{.j} = 1$

that maximize

$$\begin{aligned} E[f(X)g(Y)] &= \sum_{i=1}^{I}\sum_{j=1}^{J} a_i p_{ij} b_j \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} (a_i \sqrt{p_{i.}}) \frac{p_{ij}}{\sqrt{p_{i.}}\sqrt{p_{.j}}} \left(b_j \sqrt{p_{.j}}\right) \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} c_i q_{ij} d_j \end{aligned}$$

where

$$\begin{aligned} q_{ij} &= \frac{p_{ij}}{\sqrt{p_{i.}}\sqrt{p_{.j}}} \\ c_i &= a_i \sqrt{p_{i.}}; \quad d_j = b_j \sqrt{p_{.j}} \\ \sum_{i=1}^{I} c_i \sqrt{p_{i.}} &= 0; \quad \sum_{i=1}^{I} c_i^2 = 1 \\ \sum_{j=1}^{J} d_j \sqrt{p_{.j}} &= 0; \quad \sum_{j=1}^{J} d_j^2 = 1 \end{aligned}$$

If

$$\begin{aligned} c &= \begin{pmatrix} c_1 \\ \vdots \\ c_I \end{pmatrix} = \begin{pmatrix} \sqrt{p_{1.}} \\ \vdots \\ \sqrt{p_{I.}} \end{pmatrix} \\ d &= \begin{pmatrix} d_1 \\ \vdots \\ d_J \end{pmatrix} = \begin{pmatrix} \sqrt{p_{.1}} \\ \vdots \\ \sqrt{p_{.J}} \end{pmatrix} \end{aligned}$$

Then,

$$\begin{aligned} E[f(X)g(Y)] &= c'Qd = \sum_{i=1}^{I}\sum_{j=1}^{J} c_i q_{ij} d_j \\ &= \sum_{i=1}^{I}\sum_{j=1}^{J} \sqrt{p_{i.}} \frac{p_{ij}}{\sqrt{p_{i.}}\sqrt{p_{.j}}} \sqrt{p_{.j}} = 1 \end{aligned}$$

We require that $\sum_{i=1}^{I} c_i \sqrt{p_{i.}} = 0$, but

$$\sum_{i=1}^{I} c_i \sqrt{p_{i.}} = \sum_{i=1}^{I} \sqrt{p_{i.}}\sqrt{p_{i.}} = \sum_{i=1}^{I} p_{i.} = 1$$

Therefore, the first singular vector of Q does not work: the constraint is not satisfied. But the second singular vectors will work. If c is the second left singular vector, then

$$\sum_{i=1}^{I} c_i \sqrt{p_{i.}} = 0$$

since the first and second left singular vectors are orthogonal. Similarly with d being the second right singular vector.

The maximal correlation coefficient is then the second singular value, λ_2, and the transformed values for X and Y are given by the components of the vectors a and b, respectively.

If the second singular value is also 1, then the first and second singular vectors are not unique. Since we desire the first left singular vector to have $\sqrt{p_i}$ for its i^{th} component, we can determine a second singular vector to be orthogonal to the first singular vector and be in the span of the first two singular vectors as calculated by the SVD. Similary for the second right singular vector.

(Rényi, 1959) also proved some other properties of the maximum correlation coefficient.

$$\begin{aligned} E[f(X)|Y = y] &= \rho_{max}(X, Y)g(y) \\ E[g(Y)|X = x] &= \rho_{max}(X, Y)f(x) \end{aligned}$$

These relations are the basis for an iterative scheme for determining a nonlinear regression curve developed by (Breiman and Friedman, 1985), who called the method ACE for Alternating Conditional Expectation. Let X and Y be numerically valued random variables. Find a function $R(x)$, called the regression curve, that minimizes $E[(Y - R(X))^2]$

$$R(x) = E[Y|x]$$

Define $\| Z \| = \sqrt{E[Z^2]}$. Find functions f and g that minimize $E[(g(Y) - f(X))^2]$.

- Set $g_0(Y) = \frac{Y}{\|Y\|}$
- Iterate until there is no decrease in $E[(f(X) - g(Y))^2]$
  - $f_{n+1}(X) = E[g_n(Y) \mid X]$
  - $g_{n+1}(Y) = \frac{E[f_{n+1}(X) \mid Y]}{\|E[f_{n+1}(X) \mid Y]\|}$

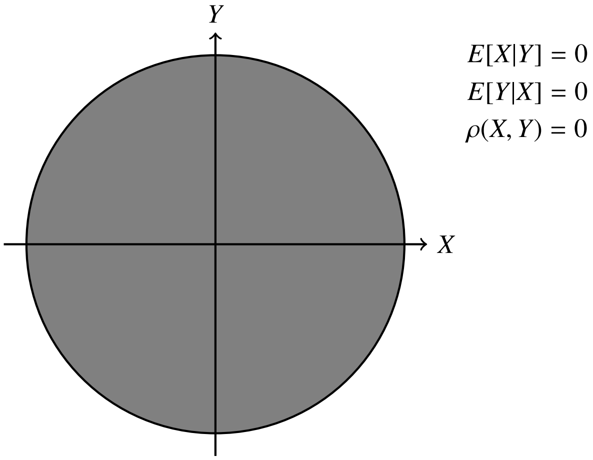


Fig. 4: Example of a uniform distribution within a unit circle.

Note that functions from F and G, minimizing $E[(f(x) - g(y))^2]$ will maximize $E[f(X)g(Y)]$ since $E[(f(X) - g(Y))^2] = 2(1 - E[f(X)g(Y)])$.

(Fowlkes and Kettenring, 1985) show the efficacy of the iterative method for data that has been sampled from the uniform distribution in a circle.

(Chernyshov, 2015) designs a method for transforming dependency measures that meet all the conditions but (C) and (G) into a measure that meets all the conditions. Feizi et al. (2015) generalize the pairwise maximal correlation coefficient into the multivariate setting by determining functions for which the sum over all pairs of transformed variables of the expected value of the product of the transformed pair is maximized. They call it Network Maximal Correlation.

Definition 3. *For random variables $X_1, \ldots, X_N$ let*

$$F_n = \{f : \mathbb{R} \to \mathbb{R} \mid E[f(X_n)] = 0 \textit{ and } V[f(X_n)] = 1\},\ n = 1, \ldots, N$$

For any graph $\mathcal{G} = (\{1, \ldots, N\}, \mathcal{E})$, the **Network Maximal Correlation** *$\rho_{\mathcal{G}}$ is defined by*

$$\rho_{\mathcal{G}}(X_1, \ldots, X_N) = \max_{f_1,\ldots,f_N : f_n \in F_n} \sum_{(i,j)\in\mathcal{E}} E[f_i(X_i)f_j(X_j)]$$

The maximum correlation has some properties that are perhaps unexpected. For example, there exist instances where

$$\begin{aligned} E[X|Y] &= 0 \\ E[Y|X] &= 0 \\ \rho(X, Y) &= 0 \end{aligned}$$

yet

$$\rho_{max}(X, Y) > 0$$

Perhaps the simplest such example is the uniform distribution within a unit circle as shown in Figure (4). Clearly, here there is no intuitive correlation. However, (Csàki and Fischer, 1963) showed that

$$\begin{aligned} \rho(X^2, Y^2) &= -\frac{1}{3} \\ \rho_{max}(X, Y) &= \frac{1}{3} \end{aligned}$$

In fact (Csàki and Fischer, 1963) proved a general case. Let $p > 0$ and (X, Y) have the uniform distribution in the set

$$\{(x, y) \mid |x|^p + |y|^p \leq 1\}$$

Then

$$\rho_{max}(X, Y) = \frac{1}{p+1}$$

(Dembo et al., 2001) constructed a more sophisticated example. Let U_1, U_2, W be independent random variables with

$$\begin{aligned} P(U_i = -1) &= 1/2,\ i = 1, 2 \\ P(U_i = 1) &= 1/2,\ i = 1, 2 \\ 0 < V[W] &< \infty \end{aligned}$$

Define

$$\begin{aligned} X_1 &= U_1 W \\ X_2 &= U_2 W \end{aligned}$$

Then, it is clear that $|X_1| = |X_2| = |W|$. And,

$$\begin{aligned} E[X_1|X_2] &= 0 \\ E[X_2|X_1] &= 0 \\ \rho(X_1, X_2) &= 0 \\ P(X_1^2 = X_2^2) &= 1 \\ \rho_{max}(X_1, X_2) &= 1 \end{aligned}$$

The principal problem with the maximal correlation coefficient as we have seen from examples where the maximal correlation coefficient is non-zero where we might expect that it should be zero is the set of functions over which the sup is taken. These examples show instances in which the optimizing function had the property that $f(x) = f(-x)$. The class of allowed functions included functions that are not one-one functions. This led (Kimeldorg and Sampson, 1978), to define the Monotone Correlation Coefficient.

Definition 4. *Let X and Y be random variables and*

$$\begin{aligned} F &= \{f : \mathbb{R} \to \mathbb{R} \mid f \textit{ is one-one}; E[f(X)] = 0; V[f(X)] = 1\} \\ G &= \{g : \mathbb{R} \to \mathbb{R} \mid g \textit{ is one-one}; E[g(Y)] = 0; V[g(Y)] = 1\} \end{aligned}$$

The **Monotone Correlation Coefficient** *$\rho_{mono}(X, Y)$ is defined by*

$$\rho_{mono}(X, Y) = \sup_{f\in F, g\in G} E[f(X), g(Y)]$$

It immediately follows from this definition that

$$\rho(X, Y) \leq \rho_{mono}(X, Y) \leq \rho_{max}(X, Y)$$

Perhaps the first question that should be asked, does independence follow if the monotone correlation coefficient is zero. The answer is affirmative and this was proved by (Kimeldorg and Sampson, 1978).

Etesami and Gohari (2016) proved the following. Let X and Y be random variables and f, g be any functions satisfying $V[f(X)] < \infty$ and $V[g(Y)] < \infty$. Then

$$\rho_{mono}(X, Y) \geq \rho_{mono}(f(X), g(Y))$$

And Etesami and Gohari (2016) also proved: Let f, g be strictly monotonically increasing functions, then for random variables X and Y,

$$\rho_{mono}(X, Y) = \rho_{mono}(f(X), g(Y))$$

Monotone correlation leads one to think about its relation to rank correlation. And indeed there is a relation: monotone correlation is never smaller than the Spearman Rank Correlation. But first for some definitions.

Definition 5. *Let $x_1, \ldots, x_N$ be N independent observations of a random Variable X. The* **Rank** *r_k of an observation x_k is defined by*

$$\begin{aligned} r_k &= |\{n \in \{1, \ldots, N\} \mid x_n < x_k\}| + \\ & \frac{1}{2}[1 + |\{n \in \{1, \ldots, N\} \mid x_n = x_k\}|] \end{aligned}$$

Suppose the ordered observations are $3.6, 5.2, 6.1, 6.1, 6.1, 7.3$. Then the ranks are $1, 2, 4, 4, 4, 6$.

Definition 6. *Let $< (X_1, Y_1), \ldots, (X_N, Y_N) >$ be a sequence of independent observations of a pairs of random variables governed by the same joint distribution function. Let $< (R_1, S_1), \ldots, (R_N, X_N) >$ be the sequence of the corresponding ranks. The* **Spearman Rank Correlation** *is the correlation of the ranks and is defined by*

$$\rho_S = \frac{\frac{1}{N}\sum_{n=1}^{N}(R_n - \bar{R})(S_n - \bar{S})}{\sigma_R \sigma_S}$$

where

$$\begin{aligned} \bar{R} &= \frac{1}{N}\sum_{n=1}^{N} R_n; \qquad \bar{S} = \frac{1}{N}\sum_{n=1}^{N} S_n \\ \sigma_R^2 &= \frac{1}{N}\sum_{n=1}^{N}(R_n - \bar{R})^2; \quad \sigma_S^2 = \frac{1}{N}\sum_{n=1}^{N}(S_n - \bar{S})^2 \end{aligned}$$

Both (Kimeldorg and Sampson, 1978) and Etesami and Gohari (2016) proved

$$\rho_{mono}(X, Y) \geq \rho_S(X, Y)$$

4. Coefficient of Intrinsic Determination

(Hsing et al., 2005) took a completely different approach to define what they called the coefficient of intrinsic determination in applications of selecting features. They had a different criteria set than Rènyi and wanted a measure that was equally applicable to continuous and categorical distributions.

(A) The measure should be model-free in the sense that no distributional or functional assumptions are placed on the variables

(B) It should be invariant under monotone transformations of the variables

(C) It can differentiate different levels of dependence. The measure of dependence of a response variable on a predictor variable should become stronger if additional information is included in the predictor variable

(D) The measure is equally applicable to continuous and categorical distributions

(E) The measure should not necessarily be symmetric

(F) The measure should be easily estimated from data

(G) The measure should be extendable to the multivariate setting

Their motivating idea is that Y should be said to be strongly dependent on X if and only if the conditional cumulative distribution of Y given X is significantly different from the marginal cumulative distribution function of Y.

Definition 7. *Let $X_1, \ldots, X_N$ and Y be random variables. The* **Coefficient of Intrinsic Dependence** *of Y given $X = (X_1, \ldots, X_N)$, $CID(Y|X)$ is defined by*

$$CID(Y|X) = \frac{\int_0^1 E_X[P(\tilde{Y} \leq u|X) - P(\tilde{Y} \leq u)]^2 du}{\int_0^1 V[P(\tilde{Y} \leq u)]du}$$

where $\tilde{Y} = F_Y(Y)$, the cumulative distribution function for Y.

5. Information Theoretic Measures of Dependency

Unlike the maximal correlation coefficient whose very definition requires variables to be numerically valued, information theoretic measures are equally applicable to numeric and categorically valued variables. This is because the definitions do not involve arithmetic operations on the values of the variables.

All the information theoretic measures involve measuring uncertainty. The uncertainty that a probability distribution has is the amount of uncertainty concerning the outcome of an experiment, the possible results of which have the probabilities $p_1, p_2, \ldots, p_N$.

Uncertainty is measured by the Shannon entropy of the distribution which is defined by:

Definition 8. *The* **Shannon Entropy** *of a random variable X taking on N possible distinct values with probabilities $p_1, \ldots, p_N$ is defined by*

$$H(X) = H(p_1, \ldots, p_N) = -\sum_{n=1}^{N} p_n \log p_n$$

By convention all log functions appearing in this paper are taken to the base 2.

Shannon's noiseless source coding theorem states that if there are N characters with probabilities $p_1, \ldots, p_N$ and a source is transmitting a sequence of characters chosen independently in accordance with probabilities $p_1, \ldots, p_N$, then there exists a code such that the average number of bits needed to encode the characters has length L where $H(p_1, \ldots, p_N) \leq L < H(p_1, \ldots, p_N) + 1$.

There are various sets of conditions that lead to the Shannon Entropy. One set is

1. $H(p_1, \ldots, p_N)$ is a symmetric function
2. $H(p, 1-p)$ is a continuous function of p
3. $H(1/2, 1/2) = 1$
4. If $p_N = q_1 + q_2 > 0$ then

$$H(p_1, \ldots, p_{N-1}, q_1, q_2) = H(p_1, \ldots, p_N) + p_N H\left(\frac{q_1}{p_N}, \frac{q_2}{p_N}\right)$$

Properties of H include

- $H(X) \geq 0$
- $H(X) \leq \log N$, where X can take N distinct values
- $H(f(X)) \leq H(X)$ for any function f

Although the Shannon Entropy is the most commonly used one, (Rényi, 1961) developed a generalization, called α-entropy.

Definition 9. *Let $\alpha > 0$ and $\alpha \neq 1$, then the α-***Entropy***, H_α, is defined by*

$$H_\alpha(p_1, \ldots, p_N) = \frac{1}{1-\alpha} \log\left(\sum_{n=1}^{N} p_n^\alpha\right)$$

The α-entropy satisfies the entropy postulates (1), (2) and (3). Furthermore,

$$\lim_{\alpha \to 1} H_\alpha(p_1, \ldots, p_N) = -\sum_{n=1}^{N} p_n \log p_n$$

(Meza et al., 2017) used α-entropy in kernel-based dimensionality reduction. (Kumar and Hooda, 2008) discuss a variety of generalized measures of entropy and dependence.

Let X take values from the set $\{\alpha_1, \ldots, \alpha_I\}$ and Y take values from the set $\{\beta_1, \ldots, \beta_J\}$. Define $p_{ij} = P(X = \alpha_i, Y = \beta_j)$. Then $H(X, Y)$ is given by

$$H(X, Y) = -\sum_{i=1}^{I}\sum_{j=1}^{J} p_{ij} \log p_{ij}$$

Some of the properties of $H(X, Y)$ include

- $H(X, Y) \geq 0$
- $H(X, Y) = H(Y, X)$
- $\max\{H(X), H(Y)\} \leq H(X, Y) \leq H(X) + H(Y)$
- $H(X, Y) = H(X) + H(Y)$ if X and Y are independent
- $H(X, Y) = H(X) = H(Y)$ if $Y = f(X)$ and f is a one-one function

Darbellay and Vajda (2000) give analytic entropy expressions for some common multivariate continuous distributions. (Nadarajah and Zografos, 2005) give expressions for Rènyi and Shannon entropies for a variety of bivariate distributions.

The mutual information, $I(X; Y)$, between two discretely valued random variables X and Y is the excess entropy of the marginal distributions over the joint distribution. It is as measure of the degree to which the joint probability distribution differs from the distribution defined by the product of the marginal probabilities. Let $p_{i.} = \sum_{j=1}^{J} p_{ij}$ and $p_{.j} = \sum_{i=1}^{I} p_{ij}$

Definition 10. *The* **Mutual Information** *between X and Y is defined by*

$$\begin{aligned} I(X; Y) &= \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{p_{i.} p_{.j}} \\ &= H(X) + H(Y) - H(X, Y) \end{aligned}$$

In coding theory, if X is the variable for the symbol being sent and Y is the variable for the symbol received, then $I(X; Y)$ is a measure of the information transmitted through the channel. For a noise-free channel, $I(X; Y) = H(X) = H(Y)$. For a channel where the symbol received is independent of the symbol sent, $I(X; Y) = 0$.

The properties of mutual information include

- $I(X; Y) \geq 0$
- $I(X; Y) = 0$ if and only if X and Y are independent
- $I(X; Y) = I(Y; X)$
- $I(f(X); g(Y)) \leq I(X, Y)$
- $I(X; X) = H(X)$
- $I(X; Y) = H(X)$ if and only if X is a function of Y
- $I(X; Y) = H(Y)$ if and only if Y is a function of X
- $I(X; f(X)) = H(X)$ for any one-one function f
- $I(X; Y) \leq \min\{H(X), H(Y)\}$
- $I(X; Y) = -\frac{1}{2}\log(1 - \rho^2)$ if X and Y are bivariate normal

(Linfoot, 1957) defined an informational coefficient of correlation.

Definition 11. *Let X and Y be random variables with mutual information $I(X; Y)$. The* **Informational Coefficient of Correlation** *is defined by*

$$\rho_{mutual}(X, Y) = \sqrt{1 - \exp(-2I(X; Y))}$$

The informational coefficient of correlation satisfies all of Rènyi's conditions with the exception of (E). It follows from this definition that if X and Y are random variables having a joint normal distribution with correlation ρ. Then,

$$\rho(X,Y) = \rho_{mutual}(X,Y)$$

(Reza, 1961) (p.146) defined a mutual information measure for a point observation $(X = u, Y = v)$.

Definition 12. *The* **Pointwise Mutual Information** *for the observation $(X = u, Y = v)$ for random variables X and Y is defined by*

$$pmi(X = u; Y = v) = \log \frac{p(X = u, Y = v)}{p(X = u)p(Y = v)}$$

The expectation of the pointwise mutual information is the mutual information.

$$E_{XY}[pmi(X;Y)] = I(X;Y)$$

Note that $pmi(X = u; Y = v)$ is positive when the joint probability is greater than what would be expected if the two variables were independent and is negative when the joint probability is less than what would be expected if the two variables were independent and is zero when X and Y are independent. (Church and Hanks, 1990) were the first ones to use pointwise mutual information in the calculation of word associations in the computational linguistics area. When $p(X = u; Y = v) = p(X = u) = P(Y = V)$, then $pmi(X = u; Y = v) = -\log P(X = u; Y = v)$ which can be very large and positive. This is not unusual to happen when the event $(X = u)$ and the event $(Y = v)$ are highly related and the marginals are small. (Thanopoulos et al., 2002) argued that the *pmi* measure needed to be normalized by subtracting from *pmi* the self information of the event $(X = u, Y = v)$ which is $-\log p(X = u, Y = v)$. He called the normalized measure mutual dependence.

Definition 13. *The* **Mutual Dependence** *of the event $(X = u, Y = V)$ is defined by*

$$md(X = u; Y = v) = \log \frac{p^2(X = u, Y = v)}{p(X = u)P(Y = v)}$$

(Bell, 1962) defined two normalized mutual information measures

Definition 14. Bell's Normalized Mutual Information *measures are defined by*

$$\begin{aligned} C'(X,Y) &= \frac{I(X;Y)}{\min\{H(X),H(Y)\}} \\ C''(X,Y) &= \frac{I(X;Y)}{\max\{H(X),H(Y)\}} \end{aligned}$$

They have the following properties:

- $0 \le C''(X,Y) \le C'(X,Y) \le 1$
- $C''(X,Y) = C'(X,Y) = 0$ if and only if X and Y are independent
- $C'(X,Y) = C'(Y,X)$
- $C''(X,Y) = C''(Y,X)$
- $C'(X,Y) = 1$ if and only if $X = f(Y)$ for some function f
- $C''(X,Y) = 1$ if and only if $X = f(Y)$ for some one-one function f
- $C''(X,Y) = 0$ implies $C'(X,Y) = 0$
- $C'(X,Y) = 0$ implies $\rho_{max}(X,Y) = 0$

He notes that no two of the three measures $\rho_{max}(X,Y)$, $C'(X,Y)$, and $C''(X,Y)$ are equivalent. Furthermore, for strictly positive probability spaces and when $H(X,Y) < \infty$, C' and C'' satisfy all seven of Rènyi's properties.

Bell makes two modifications in Rènyi's property (E).
(E′): $\delta(X,Y) = 1$ if and only if there is strict dependence $X = g(Y)$ or $Y = f(X)$
(E″): $\delta(X,Y) = 1$ if and only if X and Y are functions of each other
He also modifies (G).
(G′): $\delta(X,Y)$ is a strictly monotone function of $|\rho(X,Y)|$, if the joint distribution of X and Y is normal
Now C' satisfies (E′) and (G′); C'' satisfies (E″) and (G′). ρ_{max} does not satisfy (E′) or (E″).

(Kvalseth, 1987) argues that Bell's C' is a good measure of dependency.

There are some information theoretic forms that are metrics. (Cover and Thomas, 1991)(p.46), (Meilă, 2007), (Meilă, 2003) and (Meilă, 2005) all mention that

$$\begin{aligned} d_1(X,Y) &= 2H(X,Y) - H(X) - H(Y) \\ &= H(X,Y) - I(X;Y) \\ &= H(X) + H(Y) - 2I(X,Y) \end{aligned}$$

is a metric. Notice that with this definition, if X and Y are independent, then $d_1(X,Y) = H(X,Y) = H(X) + H(Y)$. $d_1(X,Y) = 0$ if and only if $Y = f(X)$ for some one-one function.

(Kraskov et al., 2005) proved that

$$\begin{aligned} d_1'(X,Y) &= 1 - \frac{I(X,Y)}{H(X,Y)} \\ &= \frac{2H(X,Y) - H(X) - H(Y)}{H(X,Y)} \\ d_2(X,Y) &= 1 - \frac{I(X,Y)}{\max\{H(X),H(Y)\}} \end{aligned}$$

are metric normalized to take values in the interval $[0,1]$.

(Vinh et al., 2010) proved that d_3 is a metric.

$$d_3(X,Y) = H(X,Y) - min\{H(X),H(Y)\}$$

(Horibe, 1985) proved that

$$\begin{aligned} d_4(X,Y) &= \frac{H(X,Y) - \min\{H(X),H(Y)\}}{\max\{H(X),H(Y)\}} \\ &= \begin{cases} \frac{H(X,Y)-H(Y)}{H(X)} & \text{if} \quad H(X) > H(Y) \\ \frac{H(X,Y)-H(X)}{H(Y)} & \text{otherwise} \quad H(X) \le H(Y) \end{cases} \end{aligned}$$

is a metric.

$d'_1(X,Y) = d_2(X,Y) = d_4(X,Y) = 1$ if and only if X and Y are independent and $d'_1(X,Y) = d_2(X,Y) = d_4(X,Y) = 0$ if and only if $Y = f(X)$ for some one-one function.

The forms that are metrics bring the information theoretic measures a little bit closer to measures used with variables that are numerically valued. The metrics which are normalized so that they take values in the interval $[0,1]$ are important because the similarity $s(X,Y)$ between variables X and Y can be defined by

$$s(X,Y) = 1 - d(X,Y) \tag{6}$$

where $d(X,Y)$ is one of the normalized metrics. Because of the triangular inequality satisfied by metrics, a similarity function defined by (6) inherits an interesting property:

$$s(X,Z) \ge s(X,Y) + s(Y,Z) - 1$$

(Watanabe, 1960) defined what he called the total correlation coefficient for variables $X_1,\dots,X_N$. His paper provides theorems by which the total correlation coefficient can be decomposed in terms of the partial correlations of various subsets of $\{X_1,\dots,X_N\}$.

Definition 15. *The* **Total Correlation** *among variables $X_1,\dots,X_N$ is defined by*

$$C_T(X_1,\dots,X_N) = \sum_{n=1}^{N} H(X_n) - H(X_1,\dots,X_N)$$

Let $I = \{1,\dots,N\}$ be the set of the indices for the random variables. Let variable X_n take on values in range set L_n, $n \in I$. Let $Q = \{Q_1,Q_2\}$ be a partition on I, and let $Q_1 = \{i_1,\dots,i_{K_1}\}$ and $Q_2 = \{j_1,\dots,j_{K_2}\}$ and $K_1 + K_2 = N$. We denote the projection operator projecting $x = (x_1,\dots,x_N)$ to the components indexed by $J \subset I$ by $\pi_J(x)$. Then we define[3]

$$\begin{aligned} p(\pi_{Q_1}(x)) &= p_1(x_{i_1},\dots,x_{i_{K_1}}) \\ &= \sum_{x_{j_1}\in L_{j_1},\dots,x_{j_{K_2}}\in L_{j_{K_2}}} p(x_1,\dots,x_N) \\ p(\pi_{Q_2}(x)) &= p_2(x_{j_1},\dots,x_{j_{K_2}}) \\ &= \sum_{x_{i_1}\in L_{i_1},\dots,x_{i_{K_1}}\in L_{i_{K_1}}} p(x_1,\dots,x_N) \end{aligned}$$

$$\begin{aligned} H(Q_1) &= -\sum_{u_1\in L_1,\dots,u_K\in L_{K_1}} p_1(u_1,\dots,u_{K_1}) \log p_1(u_1,\dots,u_{K_1}) \\ H(Q_2) &= -\sum_{v_1\in L_1,\dots,v_K\in L_{K_2}} p_2(v_1,\dots,v_{K_2}) \log p_2(v_1,\dots,v_{K_2}) \end{aligned}$$

Suppose that the values of the variables indexed by Q_2 have been observed. Then the ignorance about the values of the variables indexed by Q_1 becomes

$$\begin{aligned} &H(Q_1|x_j : j \in Q_2) = \\ &-\textstyle\sum_{x_i\in L_i:i\in Q_1} p(x_i : i \in Q_1 \mid x_j : j \in Q_2) \log p(x_i : i \in Q_1 \mid x_j : j \in Q_2) \end{aligned}$$

Then its expected value is

$$\begin{aligned} E[H(Q_1|x_j : j \in Q_2)] &= \sum_{x_j\in L_j:j\in\pi_2} p_2(x_j : j \in Q_2) H(Q_1|x_j : j \in Q_2) \\ &= -\sum_x p(x) \log \frac{p(x)}{p(\pi_{Q_2}(x))} \\ &= H(I) - H(Q_2) \end{aligned}$$

Before any observations, the ignorance about the variables indexed by Q_1 is $H(Q_1)$. After the observation of the values of variables indexed by Q_2, the ignorance is $H(I) - H(Q_2)$. Therefore the decrease in ignorance, which is the information about the values of the variables indexed by Q_1 given the observations of the values of variables indexed by Q_2, is

$$H(Q_1) - (H(I) - H(Q_2)) = H(Q_1) + H(Q_2) - H(I) \tag{7}$$

Because of the symmetry of (7) the decrease in ignorance, which is the information about the values of the variables indexed by Q_2 given the observations of the values of variables indexed by Q_1 is also $H(Q_1) + H(Q_2) - H(I)$.

Watanabe notes that by virtue of Gibb's theorem,

$$H(I) \le H(Q_1) + H(Q_2)$$

where equality holds if and only if X_i is independent of X_j for $i \in Q_1$ and $j \in Q_2$. Hence, $H(Q_1)+H(Q_2)-H(I) \ge 0$. This leads Watanabe to define a measure of strength of the information correlation between the variables indexed by Q_1 and Q_2.

Definition 16. *The* **Information Correlation** *between the variables indexed by Q_1 and Q_2 is given by*

$$C(I;Q_1,Q_2) = H(Q_1) + H(Q_2) - H(I)$$

5.1. Jensen-Shannon Divergence

(Lin, 1991) showed the way that the Jensen-Shannon Divergence bounds error rate for a Bayes rule.

Definition 17. *Let π_1,π_2 be prior probabilities $\pi_1,\pi_2 \ge 0$ with $\pi_1+\pi_2 = 1$. The* **Jensen-Shannon Divergence** *is defined by*

$$JS_\pi(p_1,p_2) = H(\pi_1 p_1 + \pi_2 p_2) - \pi_1 H(p_1) - \pi_2 H(p_2)$$

[3] In the first summation, we have to sum over the variables in Q_2 and in the second summation, we have to sum over the variables in Q_1

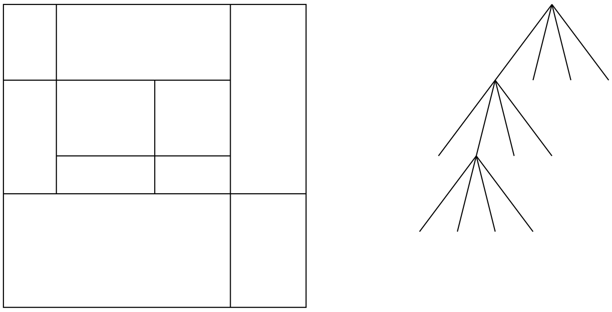


Fig. 5: Shows an example of a partition of a two-dimensional space not expressable as a product of partitions on each of the dimensions. However, its structure is a tree.

It is well known that the error rate for a Bayes rule having prior class probabilities π_1 and π_2 and class conditional probabilities p_1 and p_2 is given by

$$P_e(p_1, p_2) \quad = \quad \sum_{x \in X} \min\{\pi_1 p_1(x), \pi_2 p_2(x)\}$$

where X is the set of possible values a measurement tuple x can take.

Lin proved these lower and upper bounds on the Bayes error rate.

$$\frac{1}{4}(H(\pi_1, \pi_2) - JS_\pi(p_1, p_2))^2 \leq P_e(p_1, p_2) \leq \frac{1}{2}(H(\pi_2, \pi_2) - JS_\pi(p_1, p_2))$$

5.2. Estimating Mutual Information For Continuous Variables

There is no problem estimating the mutual information for two discretely valued variables, since everything is defined in sums. But what if the two variables are continuously valued in some interval? The natural way to do this is to partition their corresponding intervals, form the contingency table corresponding to their product partition, and then approximate the mutual information of the continuously valued variables by the mutual information of the discretized variables forming the contingency table. By a theorem of (Dobrushin, 1959), as the partitions get finer and finer the mutual information of the resulting contingency tables converge to the mutual information of the continuous variables.

One common way is to use partitions formed by equal sized intervals for each of the continuously valued variables. Another common way is to use partitions formed by intervals having equal probability. Both these ways have a problem with efficiency, the first way more than the second. Different cells in the product partition contribute to the estimate with a variable efficiency, just because some cells may have very low count and other cells have a much higher count. (Darbellay and Vajda, 1999) formulate a way to fix this problem by using an adaptive partitioning, not based on a product partition, but rather a tree partition as shown in Figure (5).

The adaptive method is as follows:

Let L_X be the range of continuous variable X and L_Y be the range of continuous variable Y. On $L_X \times L_Y$ construct a sequence of nested partitions $\pi_0, \ldots, \pi_n, \ldots$. For any cell $A \times B$ in a partition, denote by n_{AB} the number of observations that fall into the cell.

- Base Case:
  - $\pi_0 = L_X \times L_Y$
  - Define an indicator variable $c(L_X \times L_Y) = 1$
- k^{th} iteration:
  - If for all $A \times B \in \pi_k$, $c(A \times B) = 0$, then finish
  - Else
    - If $A \times B \in \pi_k$ and $n_{AB} = 0$, then put $A \times B$ in π_{k+1}
    - If $A \times B \in \pi_k$ and $c(A \times B) = 0$, then put $A \times B$ in π_{k+1}
    - If $A \times B \in \pi_k$ and $c(A \times B) = 1$, then construct a partition $\{A_1, A_2\}$ of A and a partition $\{B_1, B_2\}$ of B and let $n_{ij} = n(A_i \times B_j)$, $i, j \in \{0, 1\}$
    - Put partitions $A_i \times B_j$ in π_{k+1}, $i, j \in \{0, 1\}$
    - If the test for independence using $(n_{11}, n_{12}, n_{21}, n_{22})$ is not rejected then set $c(A, B) = 0$, and put $A \times B \in \pi_{k+1}$ else put $A_i \times B_j \in \pi_{k+1}$ and $c(A_i, B_j) = 1$, $i, j \in \{0, 1\}$

(Jain and Murthy, 2016) describe a different and fast method to estimate the mutual information. They generate a product partition in a two step procedure.

5.2.1. Testing Independence

Let n_{rc} be the number of observations in the (r, c) entry of an R by C contingency table, where X is the random variable governing the row entries and Y is the random variable governing the column entries. To test the hypothesis that X and Y are independent, the χ^2 test of (4) or the mutual information test (8) can be used.

$$2I(X;Y) \quad = \quad 2\sum_{r-1}^{R}\sum_{c=1}^{C} n_{rc} \log \frac{n_{..} n_{rc}}{n_{r.} n_{.c}} \tag{8}$$

Both the χ^2 test statistic of (4) and the mutual information test statistic of (8) have a χ^2 distribution with $(R-1)(C-1)$ degrees of freedom. In the case that $R = 2$ and $C = 2$, the test statistic is compared to $\chi^2_{1\alpha}$ where 1 is the number of degrees of freedom and α is the significance level of the test. α is the probability of the tail of the distribution for values larger than $\chi^2_{1\alpha}$.

6. Maximal Information Coefficient

(Reshef et al., 2011) state that the measure of dependence should have generality and equitability. Generality means that the measure of dependence can detect all kinds of functional relationships and not be sensitive only to linear relationships like the correlation coefficient. Equitability means that with a fixed sample size, the same amount of noise perturbing pairs of variables in different kinds of relationships should reduce the strength of the dependency in similar ways. For this purpose, (Reshef et al., 2011) define for continuous variables a maximal information coefficient.

The concept behind the maximal information coefficient of variables X and Y is to find a partition on the range of X and a partition on the range of Y such that the mutual information of the resulting product partition is the highest over all product partitions consistent with the sample size, the number of observations.

Let π_x be a partition on the range L_X of X and π_y be a partition on the range L_Y of Y. Denote the mutual information of X and Y relative to their partitions π_x and π_y by $I(X, Y; \pi_x, \pi_y)$. For $A \times B \in \pi_x \times \pi_y$, let n_{AB} denote the number of observations that fall into rectangular cell $A \times B$. Then,

$$I(X, Y; \pi_x, \pi_y) = \sum_{A \in \pi_x} \sum_{B \in \pi_y} n_{AB} \log_2 \frac{n_{AB}}{n_{A.} n_{.B}}$$

Definition 18. *The* **Maximal Information Coefficient** *between variables X and Y where a sample of N observations are made of pairs of values of X and Y is defined by*

$$MIC(X, Y; N) = \sup_{\pi_x, \pi_y} \frac{I(X, Y; \pi_x, \pi_y)}{min\{|\pi_x|, |\pi_y|\}}$$

where the sup *is taken over all pairs of partitions π_x and π_y satisfying $|\pi_x| \times |\pi_y| < N^{.6}$ and $|\pi_x|$ designated the number of cells in partition π_x.*

The algorithm used by (Reshef et al., 2011) fixed partition π_y and used a dynamic programming algorithm to find the optimal π_x. (Zhang et al., 2014) improved on this algorithm using a simulated annealing genetic algorithm. (Kinney and Atwal, 2014) challenge the properties resulting from the Reshef definition.

Finally, (Nguyen et al., 2014) defined a maximal information correlation measure for the multivariate case.

Definition 19. *Let $X_1, \ldots, X_N$ be random variables on range sets $L_1, \ldots, L_N$, respectively. For any $n \in \{1, \ldots, N\}$ let π_n be a partition on range set L_n: $\pi_n = \{I_{n1}, \ldots, I_{nm_n}\}$. Define the function $q_n : L_n \to \{1, \ldots, m_n\}$ by $q_n(x) = k$, when $x \in I_{nk}$. Define the quantized (discretized) variables $Y_n = q_n(X_n)$. Let K be a given upper bound on $|\pi_i| \times |\pi_j|$. For example $K = M^{.6}$, where M is the number of observations. The* **Maximal Information Correlation Coefficient** *ρ_{max_inf} is defined by*

$$\rho_{max\_inf}(X_1, \ldots, X_N) = \max_{\substack{\{\pi_1, \ldots, \pi_N\} \\ |\pi_i||\pi_j| < K}} \frac{\sum_{n=1}^{N} H(X_n; \pi_n) - H(X_1, \ldots, X_N; \pi_1, \ldots, \pi_N)}{\sum_{n=1}^{N} \log |\pi_n| - \max_k \log |\pi_k|}$$

(Ge et al., 2016) used the maximal information correlation measure to remove features of little association with phenotypes in a bioinformatics application.

6.1. Kullback-Liebler Divergence

For a pair of probability distributions, (Kullback, 1959) defined what has been called the Kullback I and J divergence. I is more commonly known as the Kullback-Liebler divergence, (Kullback and Liebler, 1951), although in their paper they called it the directed divergence.

Definition 20. *The Kullback I* **Divergence** *and J* **Divergence** *are defined by*

$$\begin{aligned} I(p, q) &= \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \\ J(p, q) &= I(p, q) + I(q, p) \end{aligned}$$

$I(p, q)$ is a measure of the information lost when probability distribution q is used to approximate p. It is also the amount of additional bits needed to encode independent samples coming from p using a code optimized for independent samples coming from q instead of a code optimized for independent sample coming from p.

There are many ways in which the Kullback divergence is important. One is its relationship to the L_1 difference between probability distributions.

Definition 21. *The L_1* **Difference Between Probability Distributions** *p and q defined on the same domain X is given by*

$$V(p, q) = \sum_{x \in X} |p(x) - q(x)|$$

V(p,q) is also called the **Variational Distance** *between two probability distributions.*

It follows that $0 \leq V(p, q) \leq 2$.

(Lin, 1991) proved lower bounds for I using the L_1 difference between probability distributions.

$$\begin{aligned} I(p, q) &\geq \max\{\, L_1(V(p, q)), L_2(V(p, q)) \,\} \\ L_1(v) &= \log \frac{2 + v}{2 - v} - \frac{2v}{2 + v} \qquad 0 \leq v \leq 2 \\ L_2(v) &= \frac{v^2}{2} + \frac{v^4}{36} + \frac{v^6}{288} \qquad 0 \leq v \leq 2 \end{aligned}$$

The bounds imply that as the I divergence goes to zero, the L_1 difference between two probability distributions goes to zero.

And (Lin, 1991) proved

$$\begin{aligned} K(p_1, p_2) &= \sum_{x \in X} \log \frac{p_1(x)}{1/2(p_1(x) + p_2(x))} \\ &= I(p_1, 1/2 p_1 + 1/2 p_2) \\ &\leq \frac{1}{2} I(p_1, p_2) \\ L(p_1, p_2) &= K(p_1, p_2) + K(p_2, p_1) \\ &\leq \frac{1}{2} J(p_1, p_2) \\ K(p_1, p_2) &\geq \max\left\{L_1\left(\frac{V(p_1, p_2)}{2}\right), L_2\left(\frac{V(p_1, p_2)}{2}\right)\right\} \\ L(p_1, p_2) &\leq V(p_1, p_2) \end{aligned}$$

These bounds imply that as the L_1 difference between two probability distributions goes to zero, their I divergence goes to zero.

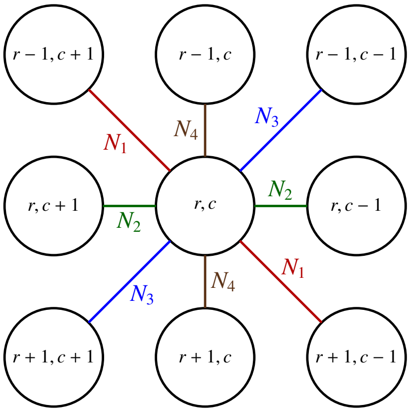
(Pinsker, 2005) proved that $\frac{1}{2} V(p_1, p_2)^2 \leq I(p_1, p_2)$. A refinement by (Fedotov et al., 2003) is the bound inequality

$$I(p_1, p_2) \leq \frac{1}{2} v^2 + \frac{1}{36} v^4 + \frac{1}{270} v^6 + \frac{221}{340200} v^8$$

where $v = V(p_1, p_2)$. (Zhang, 2007) proved a new upper bound. Let $\lambda = V(p_1, p_2)$ and $H(\lambda) = -\lambda \log \lambda - (1 - \lambda) \log(1 - \lambda)$, then

$$I(p_1, p_2) \leq 2\lambda \log(MN - 1) + H(\lambda)$$

where M is the number of of values the random variable associated with p_1 can take and N is the number of values the random variable associated with p_2 can take.



$$N_1 = \{((r,c),(u,v)) \in (R \times C)^2 \mid (u,v) = (r-1,c+1) \text{ or } (u,v) = (r+1,c-1)\}$$
$$N_2 = \{((r,c),(u,v)) \in (R \times C)^2 \mid (u,v) = (r,c+1) \text{ or } (u,v) = (r,c-1)\}$$
$$N_3 = \{((r,c),(u,v)) \in (R \times C)^2 \mid (u,v) = (r-1,c-1) \text{ or } (u,v) = (r+1,c+1)\}$$
$$N_4 = \{((r,c),(u,v)) \in (R \times C)^2 \mid (u,v) = (r-1,c) \text{ or } (u,v) = (r+1,c)\}$$

Fig. 6: The graph is the conditional independence graph of the neighborhood. Also shown is the notation for the local neighborhoods. N_1 is the lower right to upper left, N_2 is the left to right, N_3 is the lower left to upper right and N_4 is the vertical neighborhood. Next to an edge connecting two nodes is the neighborhood to which the edge belongs.

7. Texture

In this section we give a texture example visually illustrating the difference between dependency that depends on the numerical values of variables and dependency that does not depend on the numerical values of variables.

Any patch of an image that shows a texture is a region having a stochastic dependency among the pixel values of the patch. The gray level cooccurrence matrices can be used to characterize the dependency and various functionals of the cooccurrence matrix can be used as features in distinguishing one texture from another

(Haralick et al., 1973) based their textural features on the probabilities determined from the angular neighborhoods N_1, N_2, N_3, N_4 of Figure (6). For each k=1,2,3,4 they defined P_k by

$$P_k(i,j) = \frac{\#\{((r,c),(u,v)) \in N_k \mid \mathcal{I}(r,c) = i \text{ and } \mathcal{I}(u,v) = j\}}{\#N_k}$$

where $\mathcal{I}(r,c)$ is the pixel gray level at pixel location (r,c). Because of the symmetry, $P_k(i,j) = P_k(j,i)$. Let $I = J$ be the number of possible gray levels. First define

$$\begin{aligned}
P_{\{k,row\}}(i) &= \sum_{j=1}^{J} P_k(i,j) \\
P_{\{k,col\}}(j) &= \sum_{i=1}^{I} P_k(i,j) \\
\mu_k &= \sum_{i=1}^{I} iP_{\{k,row\}}(i) = \sum_{j=1}^{J} jP_{\{k,col\}}(j)
\end{aligned}$$

The features they used included:

$$\begin{aligned}
\sigma_k^2 &= \sum_{i=1}^{I} (i-\mu_k)^2 P_{\{k,row\}}(i) = \sum_{j=1}^{J} (j-\mu_k)^2 P_{\{k,col\}}(j) \\
\rho_k &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(i-\mu_k)(j-\mu_k)}{\sigma_k^2} P_k(i,j) \\
\rho &= \sum_{k=1}^{4} \rho_k \theta_k
\end{aligned}$$

where $\sum_{k=1}^{4} \theta_k = 1$ and $0 \leq \theta_k \leq 1$.

Their cooccurrence features included entropy features:

$$\begin{aligned}
E_{1k} &= \sum_{i=1}^{I}\sum_{j=1}^{J} P_k^2(i,j) \\
E_{2k} &= -\sum_{i=1}^{I}\sum_{j=1}^{J} P_k(i,j) \log P_k(i,j) \\
E_1 &= \sum_{k=1}^{K} E_{1k}\theta_k \\
E_2 &= \sum_{k=1}^{K} E_{2k}\theta_k
\end{aligned}$$

Each E_{1k} is an un-normalized form of the Rènyi generalized entropy with $\alpha = 2$. Each E_{2k} is the Shannon entropy of probability distribution P_k.

(Haralick et al., 1973) included contrast and inverse contrast features.

$$\begin{aligned}
c_k &= \sum_{i=1}^{I}\sum_{j=1}^{J} |i-j| P_k(i,j) \\
d_k &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{1}{1+\alpha|i-j|} P_k(i,j)
\end{aligned}$$

Finally, they included the maximal correlation coefficient. Let the normalized joint probability matrix $Q_k = (q_k(i,j))$ where

$$q_k(i,j) = \frac{P_k(i,j)}{\sqrt{P_{\{k,row\}}(i)P_{\{k,col\}}(j)}}$$

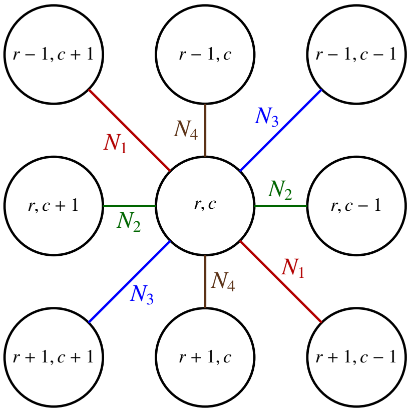
Let the second singular value of Q_k be $\lambda_{k,2}$. The maximal correlation coefficient is given by $\rho_{\{max,k\}} = \lambda_{k,2}$. [4]

(Haralick, 1975) used the cooccurrence probabilities to form a textural transform image. Let $\mathcal{I}$ be the input image, P_k, $k =$

[4] Note that in (Haralick et al., 1973), the maximal correlation coefficient was computed in a similar manner to (Witsenhausen, 1975) where $Q_k = (q_k(i,j))$ and

$$q_k(i,j) = \sum_m \frac{P_k(i,m)P_k(j,m)}{P_{\{k,row\}(i)}P_{\{k,col\}(m)}}$$

and the maximal correlation coefficient is given by $\sqrt{\lambda_2}$, where λ_2 is the second eigenvalue of the matrix Q_k.



$$\begin{aligned} P(I(u,v) : (u,v) \in N(r,c)) \quad = \quad & P(I(r,c)) \times \\ & \textstyle\prod_{(u,v)\in N_1(r,c)} P_1(I(u,v) \mid I(r,c)) \times \\ & \textstyle\prod_{(u,v)\in N_2(r,c)} P_2(I(u,v) \mid I(r,c)) \times \\ & \textstyle\prod_{(u,v)\in N_3(r,c)} P_3(I(u,v) \mid I(r,c)) \times \\ & \textstyle\prod_{(u,v)\in N_4(r,c)} P_4(I(u,v) \mid I(r,c)) \end{aligned}$$

Fig. 7: Using the conditional independence graph, this figure shows the calculation of the joint neighborhhood probability for the 3 by 3 neighborhood tree dependence.

$1,2,3,4$ be the cooccurrence probabilities, N_k, $k = 1,2,3,4$ be the local neighborhoods associated with the cooccurrence probabilities as shown in Figure (6). Define $\mathcal{J}$, the textural transform image, by

$$\mathcal{J}(r,c) \quad = \quad \sum_{k=1}^{4} \sum_{(u,v)\in N_k(r,c)} P_k(\mathcal{I}(r,c), \mathcal{I}(u,v))\theta_k$$

where the $0 \leq \theta_k \leq 1$ and $\sum_{k=1}^{4} \theta_k = 1$.

Motivated by (Haralick, 1975) we describe what can be called the joint probability image. In the joint probability image, each pixel takes the value which is the joint probability of the array of pixel gray levels in its neighborhood under a conditional independence assumption specified by the tree whose root is the center pixel. Examining the conditional independence graph, (Whittaker, 1990), of Figure (6) it can be seen that from the gray level of the center pixel, there are eight independent branches: one to the left, one to the right, one above and one below, and likewise for the four diagonal directions. Under the conditional independence graph shown in Figure (6), the joint probability for the gray levels of the of each of the neighbors given the gray value of the center pixel, can be written as the product of the conditional probabilities of each neighbors gray level given the gray level of the center pixel times the probability of the gray value of the center pixel. This is shown in Figure (7).

The idea can be extended to any kind of neighborhood, regular or non-regular. Define the 5 by 5 distance 2 neighborhood N^2 by

$$N^2(r,c) = \{(u,v) \mid (u,v) = (r,c) + (i,j),\ i,j \in \{-2,-1,0,1,2\}\}$$

For the 5×5 neighborhood, the conditional independence graph is shown in Figure (8). It can be seen that there are eight independent branches from the center pixel as is in Figure (6) to the 8 nearest neighbors of the center pixel. Then just like what happened with the center pixel and its 8 neighbors, in the 5×5 neighborhood, this happens with each of the center pixel's neighbors to their neighbors who are max_distance 2 to the center pixel. Thus the conditional independence graph of Figure (8) leads to equation (9) for the joint probability for the pixel values in the 5×5 neighborhood.

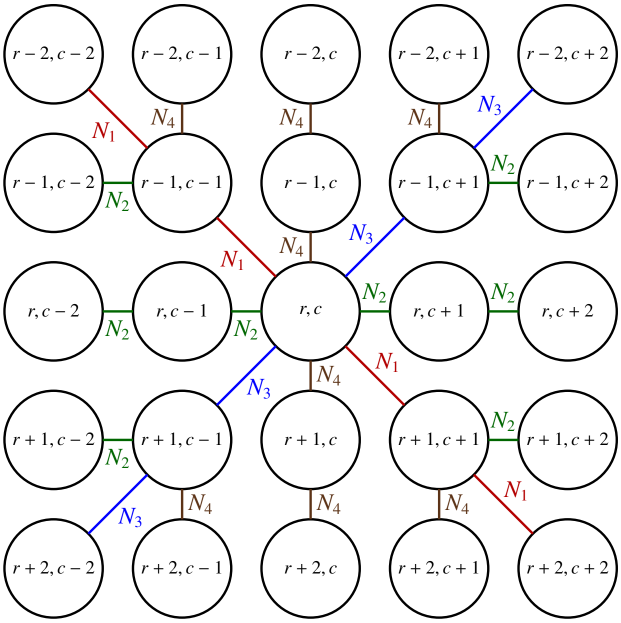
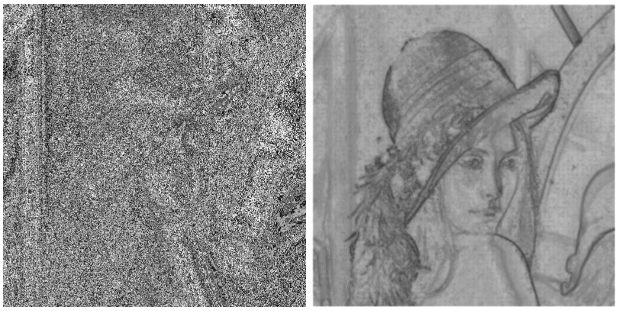


Fig. 8: Shows the joint probability tree dependence for the 5 by 5 neighborhood.

(Chow and Liu, 1968) gave an algorithm for determining the dependence tree from conditional probabilities obtained from data. Here the tree is fixed by design.

$$\begin{aligned} & P(I(u,v)(u,v) \in N^2(r,c)) = P(I(r,c)) \times \prod_{(u,v)\in N_1(r,c)} P_1(I(u,v) \mid I(r,c)) \times \\ & P_1(I(r-2,c-2) \mid I(r-1,c-1) \times P_1(I(r+2,c+2) \mid I(r+1,c+1) \times \\ & \prod_{(u,v)\in N_2(r,c)} P_2(I(u,v) \mid I(r,c)) \times P_2(I(r,c-2) \mid I(r,c-1)) \times \\ & P_2(I(r,c+2) \mid I(r,c+1)) \times P_2(I(r-1,c-2) \mid I(r-1,c-1)) \times \\ & P_2(I(r+1,c-2) \mid I(r+1,c-1)) \times P_2(I(r-1,c+2) \mid I(r-1,c+1)) \times \\ & P_2(I(r+1,c+2) \mid I(r+1,c+1)) \times \prod_{(u,v)\in N_3(r,c)} P_3(I(u,v) \mid I(r,c)) \times \\ & P_3(I(r-2,c+2) \mid I(r-1,c+1)) \times P_3(I(r+2,c-2) \mid I(r+1,c-1)) \times \\ & \prod_{(u,v)\in N_4(r,c)} P_4(I(u,v) \mid I(r,c)) \times P_4(I(r+2,c) \mid I(r+1,c)) \times \\ & P_4(I(r+2,c-1) \mid I(r+1,c-1)) \times P_4(I(r+2,c+1) \mid I(r+1,c+1)) \times \\ & P_4(I(r-2,c) \mid I(r-1,c)) \times P_4(I(r-2,c-1) \mid I(r-1,c-1)) \times \\ & P_4(I(r-2,c+1) \mid I(r-1,c+1)) \end{aligned} \tag{9}$$

Figure (9a) shows a texture image and in Figure (9b) the joint probability transform image. The calculation was done as a sum of logarithms of probabilities rather than as a product

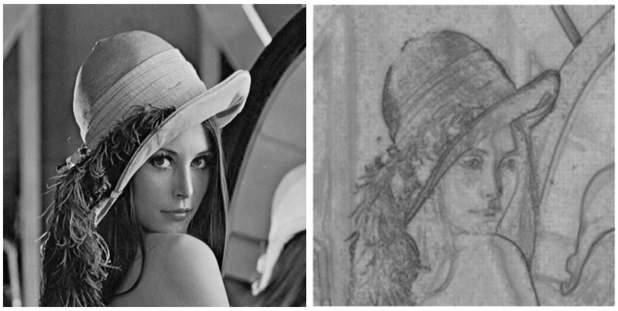


(a) Original (b) Joint Probability Transform

Fig. 9: Shows a texture image and the texture joint probability image. White means the gray level configuration in a 5x5 window in the original image has a high joint probability.

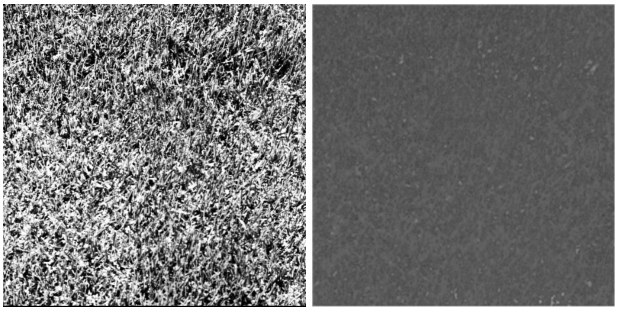
of probabilities and the resulting image was histogram equalized.[5] Everyone who has processed images will recognize the joint probability transform image as looking like some kind of edge operator operated on the Lena image as shown in Figure (10). The original image shown in Figure (9a) is actually a random gray level permutation of the Lena image of Figure (10a). Thus the spatial dependence of the gray level permuted image and the original image determined by the cooccurrence probabilities are identical. There was no edge operator based on numerical differences of gray values. The darker local areas correspond to areas having lower joint probability in their 5×5 neighborhoods. This is an example showing that the 5×5 neighborhood joint probability in a gray level randomly permuted image as shown in Figure (9a) carries the structure of the cooccurrence probabilities but does not carry the structure of contrast changes that would be indicated by statistically significant changes in gray levels across a boundary, a structure that our visual system would interpret as boundaries. Nevertheless, since the gray level arrays of local neighborhoods that have edges also have low neighborhood joint probability, the neighborhood joint probability can detect edges.

Since Figure (10) illustrates the dependency only coming from the cooccurrence probabilities and none coming from the contrast, the gray level differences across a boundary, it is probably a surprise that so much of boundary/contrast information is contained in cooccurrence probabilities. This suggests that edge operators that are based on the computation of gray level differences to estimate derivatives of the underlying sampled gray level surface should be complemented with some kind of local joint probability estimate determined by cooccurrence probabilities.



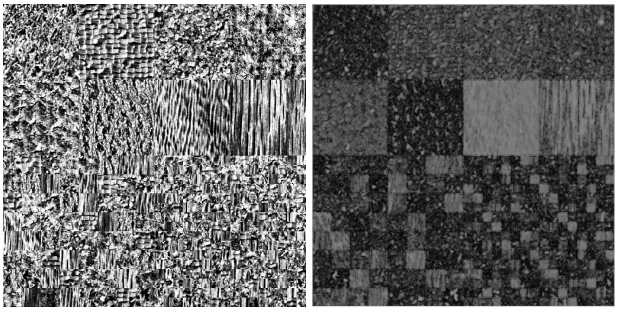
(a) Lena (b) Joint Probability Transform

Fig. 10: Lena and its joint probability transform.



(a) Original (b) Joint Probability Transform

Fig. 11: Shows one of the more uniform texture Brodatz images.



(a) Texture Mosaic (b) Joint Probability Transform

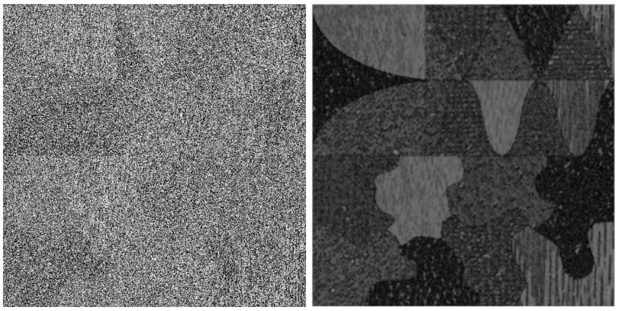
Fig. 12: Shows one of the multiple texture mosaic Brodatz images.

When the scale of the texture and size of window are comparable, the joint probability is about the same for each 5×5 window. This is shown in Figure (11). Figure (12) shows an example of a Brodatz texture mosaic. Within each uniform texture area, the joint probability in a 5×5 window is nearly the same when the texture scale and the window size are similar.

Notice that the joint probability images of Figures (12) and (13) would be suitable for an image segmentation operator, whereas the original texture mosaic would have to be processed by a segmentation operator that was designed for textures.

Figure (13) Shows another example of a texture mosaic image whose gray levels have been randomly permuted and the joint probability texture transform image.

[5] Thanks to Vishal Bharti who prepared the images and their transforms.



(a) Gray Level Permuted (b) Joint Probability Transform

Fig. 13: Shows another one of the multiple texture mosaic Brodatz images. The original image(a) is operated on with a random gray level permutation. The joint probability transform image shows the structure of the textures that were put together to make up the mosaic.

8. Manifold Methods

If there is an ideal dependency among variables, the values of the variables are constrained. This constraint can be utilized in both in the clustering context such as (Haralick and Harpaz, 2005), (Haralick and Harpaz, 2007), (Harpaz and Haralick, 2007), and (Haralick et al., 2016) and in the supervised learning context.

Suppose there are N features. If for any class c, there are subsets I_{kc}, $k = 1, \ldots, K_c$ of indexes of features that have a dependencies of the form $f_{kc}(\pi_{I_{kc}}(x)) = 0$, $k = 1, \ldots, K_c$. we can think of each of these dependencies as defining a manifold

$$M_{kc} = \{x \mid f_{kc}(\pi_{I_{kc}}(x)) = 0\}$$

Since there may be multiple such dependencies for different subsets of variables, the entire constraint set is given by

$$M_c = \bigcap_{k=1}^{K_c} M_{kc}$$

Here, M_c is the manifold containing all measurement space tuples $(x_1, \ldots, x_N)$ labeled as class c satisfying all the constraints.

This way of representing dependencies is similar to what is known as universal approximation functions, (Cybenko, 1989), (Hornik et al., 1989) and (Baron, 1993). The theorem of universal approximation states: Let f be some continuous function defined on the N-dimensional unit hypercube. Let $\varphi : \mathbb{R} \to \mathbb{R}$ be a nonconstant, bounded, and monotonically-increasing continuous function. For every $\epsilon > 0$, there exists an integer $J = J(\epsilon)$, real constants $v_j, b_j \in \mathbb{R}$ and real vectors $w_j \in \mathbb{R}^N$, satisfying

$$\left| \sum_{j=1}^{J} v_j \varphi\left(w'_j x + b_j\right) - f(x) \right| \leq \epsilon$$

Here the form $w'_j x + b_j$ can be understood as producing the projection of x onto a one-dimensional linear manifold, representing the projection relative to the coordinate system of the linear manifold. The function φ operating on that projection is a simple bounded non-linear monotonically increasing operator. The sum combines the resulting values.

The universal approximation theorem is what gives multiple layer neural networks the potential for being universal approximators.

Analogous to the universal approximation theorem, we conjecture the following kind of theorem that would make classifiers based on sub-manifolds the way to think about classifiers. Let C be the set of classes and $Q = [0, 1]^N$ be the N dimensional unit hypercube. Let $f : Q \to C$ be the desired classification function. If f is sufficiently simple, then for every $\epsilon > 0$, there exists a $J = J(\epsilon)$ and very simple functions $h_{jc} : [0, 1]^{K_c} \to \mathbb{R}$, $j = 1, \ldots, J$, $c \in C$ and $K_c \ll N$, such that

$$P\left(\{x \in Q \mid \min_{j\in J} h_{jc}(\pi_{jc}(x)) \geq \min_{j\in J} h_{jd}(\pi_{jd}(x)) \forall d \in C\} \Delta f(x)\right) \leq \epsilon$$

where each $\pi_{jc}(x)$ is an orthogonal projection operator onto some small subset of the components of x, Δ means the symmetric set difference and for any $S \subset Q$, $P(S)$ is the probability that an observation lies in the set S. Alternative versions replace the min with $\sum$ or max

$$P\left(\{x \in Q \mid \sum_{j\in J} h_{jc}(\pi_{jc}(x)) \geq \sum_{j\in J} h_{jd}(\pi_{jd}(x)) \forall d \in C\} \Delta f(x)\right) \leq \epsilon$$

or

$$P\left(\{x \in Q \mid \max_{j\in J} h_{jc}(\pi_{jc}(x)) \geq \max_{j\in J} h_{jd}(\pi_{jd}(x)) \forall d \in C\} \Delta f(x)\right) \leq \epsilon$$

If the conjecture is true, it hints that simple classification functions imply that there are dependencies among the subspaces of the features and the classes.

There are many papers on manifold learning and manifold learning with respect to applications in computer vision and signal processing. We just reference a few of them as we are not making a review of the area. See for example (Kim et al., 2015), (Liu et al., 2004), (Turaga et al., 2011), (Srivastava and Klassen, 2004) and (Lui, 2012). Brahma et al. (2016) reinforce the idea that the deep learning with feedfoward neural networks can be understood through the concept of manifolds and the way that successive layers of the networks act is to flatten the manifolds.

8.1. Subspace Classifiers and Subspace Ensemble Classifiers

A classifier characterizes the dependency between the measurement tuple and the class. In effect the classifier assigns a measurement tuple to the class having the largest positive dependency to it.

A subspace classifier is one which, for each class, projects the measurement tuple x to one or more subspaces, each of which produces a value and which for each class combines the results and assigns x to that class having the highest combined result. When the subspaces are chosen in some random way, the classifier is called an ensemble classifier. The main issues for the ensemble classifier are (1) how to generate multiple nearly independent classifiers and (2) how to combine the outputs of the classifier to make a class assignment.

The reason for the nearly independent requirement goes back to the eighteenth century Condorcet's jury theorem which states, relative to our classifier context, that if there are L independent classifiers each with probability $p > .5$ of being correct, then the probability q of the majority of the classifiers being correct must satisfy $q > p$. That is, more classifiers will lead to higher classification accuracy. And in the limit as the number of classifiers grow, q approaches 1.[6]

There are many papers on ensemble classifiers and we only give a few references since it is not our purpose here to make a comprehensive review. (Haralick, 1976) discussed how to combine results using class conditional probabilities of multiple shallow probability-based classifiers. (Kittler et al., 1998), discussed a variety of general methods for combining classifiers. (Breiman, 1996) introduced the bagging methods. (Freund and Shapire, 1997) introduced the boosting methods. (Kittler and Roli, 2000) organized one the early workshops that encompassed ensemble classifiers and other ways in which multiple classifiers can be utilized to give higher accuracy than any of the highly trained single classifiers. Reviews of ensemble classifiers can be found in papers in (Kittler and Roli, 2000),(Dietterich, 2000), (Valentini and Masulli, 2002), (Rokach, 2010) and (Džeroski et al., 2000). (Yu et al., 2016) discusses a method for ensemble learning that simultaneously uses the data sampling space and the measurement or feature space.

Here we establish the notation we use to describe any subspace ensemble classifier. Let there be N components to the measurement tuple and let $I = \{1, \ldots, N\}$ be the index set for these N components. Let there be M randomly selected subspaces and K classes. Let us denote by T_{mk} the m^{th} shallow classifier for class k. Let $Q_m \subset I$ denote the indexes specifying the randomly selected subspace for the T_{mk} shallow classifier. We denote the projection operator to this m^{th} selected subspace by π_{Q_m} Then $T_{mk}(\pi_{Q_m}(x))$ represents the strength of the dependency of the measurement tuple x with respect to the m^{th} subspace to class k as specified by the shallow classifier T_{mk}. The measurement tuple x is then assigned to class k where

$$k = \underset{i}{\operatorname{argmax}} \sum_{m=1}^{M} T_{mi}(\pi_{Q_m}(x))$$

An alternative way of combining is given by

$$k = \underset{i}{\operatorname{argmax}} \min_{M=1}^{M} T_{mi}(\pi_{Q_m}(x))$$

We begin with the N-tuple classifier, which was developed for recognizing hand printed block characters and typewritten characters by (Bledsoe and Browning, 1959). Then we discuss the subspace classifier due to (Watanabe, 1969), who developed a methodology for selecting different subspaces for different classes.

(Ho, 1995) and (Ho, 1998) was the first one to discuss the random forest method. Here the forest consists of many shallow decision trees, each one using a random subspace sampling, whose outputs are combined. (Dietterich, 1998) compared decision tree methods like bagging, boosting, and random subspace sampling. (Breiman, 2001) showed that the generalization error converges asymptotically to a limit as the number of trees in the forest becomes large.

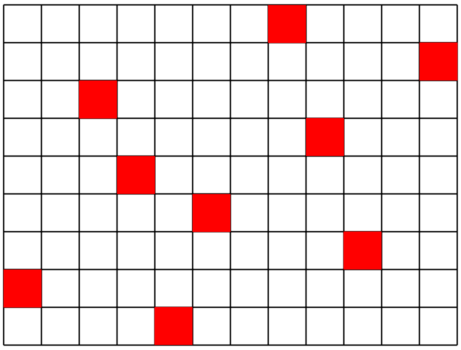


Fig. 14: Shows an example window in which the character is located. The pixels shown in red are the pixels that are selected by one of the multiple tables maintained by the N-tuple method for the classification of the printed character.

On a completely different line of thinking we have the area of Bayesian Networks, developed by (Pearl, 1988), and graphical models, developed by (Lauritzen, 1996). Neither of these is considered as a subspace classifier, but in fact each is. The method of combining is the multiplication of class conditional probabilities and all the components of the measurement tuple involved in any one conditional probability is a projection of the full measurement tuple to a subspace determined by the selected components. It is this correspondence that will allow us to suggest an alternative way to work with the N-tuple classifier.

8.2. The N-tuple Method

The N-tuple method was one of the early successes in printed character recognition over a half century ago. The method included segmenting each character into a fixed window of $M \times N$ pixels. Each pixel is thresholded so that its value is just 0 or 1. The technique was designed for specialized fast table lookup hardware and there were a variety of different character recognition hardware implementations that incorporated the N-tuple method such as the Wisard hardware, (Aleksander and Morton, 1995). Through the 1990's, many of the IBM products that had classifiers used N-tuple classifiers.

Since the Bledsoe and Browning paper, there have been many papers describing specialized hardware, experimental results, and variations of the original N-tuple classifier. (Aleksander and T.J.Stonham, 1979) give a general review. (Allinson and Kolcz, 1997) discuss how it can be used for estimating a generalized regression function. The technique has been extended to a scanning mode by (Lucas and Amiri, 1995). (Rohwer, 1995) gives a Bayesian treatment of the N-tuple method. (Rohwer and Morciniec, 1998) and (Jorgensen and Linneberg, 1999) give a theoretical analysis. There are reviews such as (Ludermir et al., 1999). And of course there are papers describing experimental results such as Morciniec and Rohwer (1995) and (Ghazanfar and Ghani-Haider, 2016), just to cite a few.

In the original method, a small number of pixel positions are randomly selected multiple times. Since the sequence of pixels values can be considered a tuple, each random selection of

[6] See https://en.wikipedia.org/wiki/Condorcet's_jury_theorem

pixels corresponds to a random sampling of a subspace.

Because of the thresholding, each of these positions has a binary 0 or a binary 1 value. The binary values of the selected positions are concatenated to form a binary number. This number is used to form an address for one of the many tables in memory. There is a set of tables for each character class and there are multiple sets of such randomly selected pixel positions for each character class.

Let there be M pattern sets of randomly selected pixel positions and K character classes. Let us denote by T_{mk} the lookup table for pattern set m and class k. Because there are M pattern sets of randomly selected pixel positions, a printed character produces M binary addresses $b_1, \ldots, b_M$. $T_{mk}(b_m)$ holds a binary 1 if some character in the training set of class c_k has the binary number b_m for the m^{th} pattern set The table lookup calculation is one of the following.

$$f_k = \min_{m=1}^{M} T_{mk}(b_m) \qquad (10)$$

$$f_k = \sum_{m=1}^{M} T_{mk}(b_m) \qquad (11)$$

The N-tuple method assigns the character to unique class c_k, if there is one, for which $f_k > 0$ is highest. Otherwise reserve decision. (Aleksander and T.J.Stonham, 1979)

In the original N-tuple classifier, the subspaces are chosen at random and the value placed in the table $T_{mk}(x)$ is related to the class k conditional probability of $\pi_{Q_m}(x)$. An improvement can be made by initializing each of the Q_m at random and then iteratively altering one of them at a time by removing one index and adding in another, keeping the alteration if the probability of correct assignment improves. Improvement can also be made by choosing a table at random and iteratively changing the value of one entry in the table so that the probability of correct assignment improves. The two searches can be done alternately. First select a Q_m and make a change to increase probability of correct assignment and then select an entry in a table and make a change to increase probability of correct assignment.

In an implementation variation, the contents of $T_{mk}(b_m)$ is the training set estimated class conditional probability $P(b_m \mid c_k)$. Then taking the minimum for each k amounts to the computation

$$\min_{m=1}^{M} P(b_m \mid c_k) \qquad (12)$$

It is easy to change the maximum likelihood form of equation (12) to a Bayesian form by incorporating the class priors. Just use

$$P(c_k) \min_{m=1}^{M} P(b_m \mid c_k)$$

In effect, the class prior probability weights the combined output for the classifier for c_k.

To explain what is the meaning of this minimum, we need some notation. Let x be the observed measurement tuple to be assigned a class. Let $I = \{1, 2, \ldots, N\}$ be an index set for the components of the measurement tuple x. Then our observed measurement tuple can be denoted by the pair (I, x). Let $Q_1, \ldots, Q_M$ be the M pattern sets. Each $Q_m \subset I$. They are designed so that $Q_1, \ldots, Q_M$ is a cover for I. $b_m(I, x)$ is essentially the linear address for the multidimensional array storing the probabilities of the projection $\pi_{Q_m}(I, x)$. This is the projection of x onto its components indexed by Q_m. Thus $P(b_m(I, x) \mid c_k) = P(\pi_{Q_m}(I, x) \mid c_k)$. Since $\cup_{m=1}^{M} Q_m = I$, the event that (I, x) arises from class c_k is the same event as $\cap_{m=1}^{M} \pi_{Q_m}(I, x)$ arises from class c_k. But for any events A and B, $P(A \cap B) \leq P(A)$ and $P(A \cap B) \leq P(B)$ so that $P(A \cap B) \leq min\{P(A), P(B)\}$. In general, $P(\cap_{m=1}^{M} A_m) \leq \min_{m=1}^{M} P(A_m)$. Therefore,

$$P((I, x) \mid c_k) = P(\cap_{m-1}^{M} \pi_{Q_m}(I, x) \mid c_k) \leq \min_{m=1}^{M} P(\pi_{Q_m}(I, x) \mid c_k)$$

Since $\min_{m=1}^{M} P(b_m(x) \mid c_k) = min_{m=1}^{M} P(\pi_{Q_m}(I, x) \mid c_k)$, assigning x to the class c_k where

$$\min_{m=1}^{M} P(b_m(x) \mid c_k) \geq \min_{m=1}^{M} P(b_m(x) \mid c_j)$$

means assigning x to the class c_k having the largest upper bound on the class conditional probabilities of the subspaces.

Let us describe another variation of the N-tuple method. As before, there are M pattern sets of randomly selected pixel positions and K character classes. A printed character x produces M binary addresses $b_1(x), \ldots, b_M(x)$. We designate by T_m the lookup table for pattern set m. $T_m(b_m)$ holds the set of classes associated with the binary address b_m for the m^{th} pattern set. A class $c \in T_m(b_m)$ if $P(c \mid b_m) > \theta_c$ Define $F = \cap_{m=1}^{M} T_m(b_m)$. The character is assigned to unique class c_k, if there is one, where $c_k \in F$ and $|F| = 1$. Otherwise reserve decision.

To explain what this variation of the N-tuple method does, we make a shift of notation to be able to discuss relations. Each of the selected pixel positions of Figure (14) is considered as a variable. Let $X_1, \ldots, X_N$ be the N variables. Let L_n be the possible values variable X_n can take. Let R be the relation containing all the tuples in the training set for one class. Since we assume that N is large, we expect that each training observation is unique. There are no duplicates. Thus,

$$R \subseteq \bigtimes_{n=1}^{N} L_n$$

Since we need to discuss various kinds of projections of a relation, we need to keep track of the what variables are associated with some N-tuple that has a selection of components of the measurement tuples of the training set. We will do this by an index set.

Definition 22. *If I is an index set and $R \subseteq \bigtimes_{i \in I} L_i$, then we say (I, R) is an* **Indexed N-ary Relation** *on the range sets indexed by I.*

Our explanation involves the operation of relation join. This is the equijoin or natural join in the database world.

Definition 23. *Let I, J, K be index sets with $K = I \cup J$. Let $R \subset \bigtimes_{i \in I} L_i$ and $\bigtimes_{j \in J} L_j$. Then the* **Relation Join** *of (I, R) with (J, S) is denoted by $(I, R) \otimes (J, S) = (K, T)$ where*

$$T = \{t \in \bigtimes_{k\in K} L_k \quad | \quad \pi_I(K,t) \in (I,R)\ \mathit{and}\ \pi_J(K,t) \in (J,S)\}$$

and $\pi_J(K,t)$ designates the projection of the tuple (K,t) onto those variables or components specified by the index set J. Likewise for $\pi_I(K,t)$.

Now we can re-express what this variation of the N-tuple method does. Let Λ_c be the set of measurement tuples that are assigned to a class c. Λ_c is the subset of measurement tuples in the relation join of the tables associated with class c.

$$\begin{aligned} \Lambda_c &= \{([N],x) \mid \pi_{J_m}([N],x) \in (J_m, T_{mc}), m = 1,\ldots,M\} \\ &= \otimes_{m=1}^{M}(J_m, T_{mc}) \end{aligned}$$

The *Acceptance Region* A_c for a class c is the set of all measurement N-tuples that will be assigned to the class.

$$A_c = \otimes_{m=1}^{M}(J_m, T_{mc}) - \bigcup_{\{d\in C-\{c\}\}} \otimes_{m=1}^{M}(J_m, T_{md})$$

The *Reserve Decision Region* is the set consisting of measurement N-tuples that do not belong to any acceptance region.

$$R = \bigtimes_{n=1}^{N} L_n - \bigcup_{c\in C} A_c$$

(Tattersall et al., 1991) suggest that the N-tuple method, which they call a single layer lookup perceptron, is in some sense an interpolation system, and thus an approximator, that interpolates, what we would say, are the class conditional probabilities from sparse training sets. (Kolcz and Allinson, 1996) argue that the N-tuple network operates as a non-parametric kernel regression estimator with the advantage that instead of having to store all the training vectors explicitly, it stores them implicitly as the sampled N-tuples and thereby uses a fixed memory size regardless of training set size. It is well known that multilayer feedfoward neural networks can function as approximators. In fact, (Hornik, 1991) proves that if sufficiently many hidden units are available, and the activation function is bounded, continuous and not constant, then continuous mappings can be learned uniformly over compact input sets.

The way that we are asking the approximator question is not with respect to class conditional probability estimation. We are posing the problem from the point of view of the classification function, which is a function from a high dimensional discrete measurement space into the small set of classes. The set of classes can be from two classes to hundreds of classes. Because of the ordering of the values in each dimension of the measurement space, it should be possible to assign a complexity measure to the classification function. Of course there is a complexity of the N-tuple memory which can be measured as the number of memory locations. Our conjecture is that there might be an approximation theorem that states that if the complexity of the classification function is less than C, then for a fixed memory size M, it is possible to approximate the desired classification function to within ϵ, where ϵ is the ratio of the number of classification differences to the size of measurement space.

8.3. The Watanabe Subspace Classifiers

The subspace classifier was introduced by Watanabe, (Watanabe, 1969). He was motivated by entropy. The entropy H of a $K - \mathit{dimensional}$ multivariate Gaussian of uncorrelated variables is given by

$$H = \frac{1}{2}(1 + log2\pi) + \frac{K}{2}\sum_{k=1}^{K} log\sigma_k$$

where σ_k is the standard deviation of the k^{th} variable.

From observations with his own data sets, he observed that for any class, the few largest eigenvalues accounted for 90% to 95% of the entropy. He based his CLAFIC (Class Featuring Information Compression) method on this idea, (Watanabe, 1970). Suppose there are M classes and there are L_m D-dimensional feature vectors $x_1^m, \ldots, x_{L_m}^m$ from class c_m. Let $N = \sum_{m=1}^{M} L_m$ be the total number of feature vectors in the training set. Define the global training set mean by μ.

$$\mu = \frac{1}{N}\sum_{m=1}^{M}\sum_{k=1}^{L_m} x_k^m$$

Define the scatter matrix for class c_m by

$$\Xi_m = \frac{1}{L_m}\sum_{k=1}^{L_m}(x_k^m - \mu)(x_k^m - \mu)'$$

Note that the mean μ is the global mean and not the class conditional means.

Order the eigenvalues of Ξ_m, $\lambda_1^m \geq \lambda_2^m \geq \ldots \geq \lambda_D^m$. Let $t_1^m, \ldots, t_D^m$ be the corresponding eigenvectors.

Given σ, $0 < \sigma < 1$, the J_m most important directions for class m are

$$t_1^m, \ldots, t_{J_m}^m$$

where

$$\frac{\sum_{j=1}^{J_m-1}\lambda_j^m}{\sum_{j=1}^{D}\lambda_j^m} < \sigma \leq \frac{\sum_{j=1}^{J_m}\lambda_j^m}{\sum_{j=1}^{D}\lambda_j^m}$$

The CLAFIC method then assigns x to class c_m where

$$\sum_{j=1}^{J_m}\left((t_j^m)'x\right)^2 \geq \sum_{j=1}^{J_k}\left((t_j^k)'x\right)^2, k = 1,\ldots,M$$

Let T_m be a matrix whose columns are the J_m orthonormal eigenvectors. Define

$$\begin{aligned} T^m &= \begin{pmatrix} \vdots & \vdots & \ldots & \vdots \\ t_1^m & t_2^m & \ldots & t_{J_m}^m \\ \vdots & \vdots & \ldots & \vdots \end{pmatrix} \\ P_m &= T^m(T^m)' \end{aligned}$$

Then P_m is the orthogonal projection operator onto the subspace spanned by $Col(T^m)$. Re-expressed in terms of the orthogonal projection operator, the CLAFIC method assigns x to class c_m where

$$||P_m x||^2 \geq ||P_j x||^2, j = 1,\ldots,M$$

This is equivalent to assign x to class c_m where

$$x'P_m x \geq x'P_j x, j = 1, \ldots, M$$

Specialized for the two class case, for any non-negative threshold θ that balances the class misdetect rates between two classes, the CLAFIC method assigns x to class c_1 if

$$\frac{x'P_1 x}{x'P_2 x} > \theta$$

Else assign x to class c_2

There is a geometric interpretation to what CLAFIC does. It assigns a vector x to that class where the angle between x and the class subspace is minimized.

Let P be an orthogonal projection operator to a subspace V
Let θ be the angle between x and V
Then

$$cos^2\theta = \frac{x'Px}{x'x}$$

Hence, assign x to class c_m when

$$x'P_m x \geq x'P_j x, \; j = 1, \ldots, M$$

is the equivalent to Assign x to class c_m when

$$\begin{array}{rcl} \frac{x'P_m x}{x'x} & \geq & \frac{x'P_j x}{x'x}, \; j = 1, \ldots, M \\ cos^2\theta_m & \geq & cos^2\theta_j, \; j = 1, \ldots, M \\ \theta_m & \leq & \theta_j, \; j = 1, \ldots, M \end{array}$$

(Watanabe and Pakvasa, 1973) further develop the subspace method. They argue that among the M classes whose subspaces are $S_1, \ldots, S_M$, there might be a non-trivial common subspace. Such a subspace should not play a role in the classification. Therefore, that subspace should be subtracted out of the projection operator.

(Therrien, 1975) details an easy way to do this. He proves: Let $P_m, m = 1, \ldots, M$ be the orthogonal projection operators to subspaces $S_1, \ldots, S_M$. Let $S = \cap_{m=1}^{M} S_m$. Let $\Gamma = \sum_{m=1}^{M} a_m P_m$ where the a_m's are chosen so that $0 < a_m < 1$ and $\sum_{m=1}^{M} a_m = 1$. Then the orthogonal projection operator Q onto S is given by $Q = TT'$, where the columns of T are the eigenvectors of Γ having eigenvalue 1.

To implement Watanabe's idea, if $P_1, \ldots, P_M$ are the orthgonal projection operators to the subspaces $S_1, \ldots, S_M$ and if Q is the orthogonal projection to the common subspace $S = \cap_{m=1}^{M} S_m$, then the projection operators that project to their subspaces minus their common subspaces are $Q_1, \ldots, Q_M$ where $Q_m = P_m - Q$. This is true because $P_m Q = QP_m = Q$ which makes $P_m - Q$ be the orthogonal projection operator to the subspace $S_m \cap (\cap_{i=1}^{M} S_i)^{\perp}$. As before, assign x to that class c_k where $x'Q_k x \geq x'Q_m x, m = 1, \ldots, M$.

Since the 1973 Watanabe paper, there have been many others on subspace methods. We just highlight a few. (Oja, 1983) wrote a book all about subspace methods. (Prakash and Murty, 1996) clustered the training vectors for each class first and then used the method of (Watanabe and Pakvasa, 1973) to identify the best cluster and then the best class. For any vector x find best cluster among all the clusters of all the classes and assign x to the class associated with that best cluster. (Watanabe and Katagiri, 1995) (not Satosi Watanabe) reframed the problem. Instead of using the subspaces which best fit the training vectors from each class, they changed the criteria: use the subspaces that are most discriminative; the Minimum Error Learning Subspace. Their procedure is an iterative gradient search procedure.

(Yin et al., 2014) and (Liang et al., 2016) are example papers that apply subspace learning for dimensionality reduction. Nikitidis et al. (2014) brings in the idea of the maximum margins for determining the subspaces. It is interesting that many of the recent papers on subspace classifiers do not reference their origin in the Watanabe papers from the 1970's.

8.4. Graphical Models as Subspace Classifiers

There are many papers and some books, (Lauritzen, 1996) and (Koller and Friedman, 2009), to cite a couple, on graphical models. However it is not the purpose of this section to make a review of graphical models. Our purpose is just to show that when a graphical model is used to express the class conditional probabilities, then the associated classifier is a subspace classifier. And from this relation, there follows a way to optimize ensemble classifiers.

Graphical Models associate a graph, called the conditional independence graph, from which all the conditional independencies can be easily seen.

Definition 24. *A graph $G = (V, E)$ is called a* **Conditional Independence Graph** *of a random variable set $\Lambda = \{X_1, \ldots, X_M\}$ if and only if $V = \{1, \ldots, M\}$, the index set for the variables in Λ, and*

$$E^c = \{\{i, j\} \mid \{i\} \perp\!\!\!\perp \{j\} \mid \Lambda - \{i, j\}\}$$

where $A \perp\!\!\!\perp B \mid C$ means the variables indexed in A are conditionally independent of the variables indexed in B given the variables in C.

Definition 25. *A graph is called* **Triangulated (Chordal)** *graph if and only if every cycle of length 4 or more has a chord.*

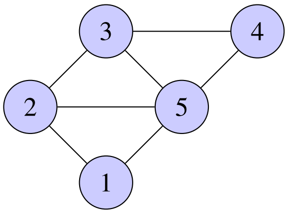
When the conditional independence graph has the property of being triangulated, then the joint probability function can be expressed with a probability product form. The product form is a strong extension of the marginal probability terms of the product and is based on the cliques and separators of the conditional independence graph.

Definition 26. *The cliques $C_1, \ldots, C_K$ of G are said to be in* **Running Intersection Order** *with separators $S_2, \ldots, S_K$ if and only if*

$$S_k = C_k \bigcap \left(\bigcup_{i=1}^{k-1} C_i \right), k = 2, \ldots, K - 1$$

and each S_k consists of the vertices of a complete graph.

C_1	$=$	$\{1,2,5\}$	$1 \perp\!\!\!\perp 4 \mid 2,5$
C_2	$=$	$\{2,3,5\}$	$1 \perp\!\!\!\perp 3 \mid 2,5$
C_3	$=$	$\{3,4,5\}$	$2 \perp\!\!\!\perp 4 \mid 3,5$
S_2	$=$	$\{2,5\}$	$1 \perp\!\!\!\perp 4 \mid 3,5$
S_3	$=$	$\{3,5\}$	$1 \perp\!\!\!\perp 4 \mid 2,3,5$



$$\begin{aligned} P(x) &= \frac{P(\pi_{C_1}(x))P(\pi_{C_2}(x))P(\pi_{C_3}(x))}{P(\pi_{S_2}(x))P(\pi_{S_3}(x))} \\ &= P(\pi_{C_1}(x))P(\pi_{C_2-S_2}(x) \mid \pi_{S_2}(x))P(\pi_{C_3-S_3}(x) \mid \pi_{S_3}(x)) \end{aligned}$$

Fig. 15: Shows a triangulated conditional independence graph for the variables indexed by $1,2,3,4,5$, the cliques and separators, some of the conditional independences and the resulting product form for the joint probability.

Proposition 1. *If a graph G is triangulated graph and $C_1,\ldots,C_K$ are the cliques of G put in running intersection order with separators $S_2,\ldots,S_K$, where*

$$S_k = C_k \bigcap \left(\bigcup_{i=1}^{k-1} C_i \right), k = 2,\ldots,K$$

then

$$\begin{aligned} P(x) &= \frac{\prod_{k=1}^K P(\pi_{C_k}(x))}{\prod_{k=2}^K P(\pi_{S_k}(x))} \\ &= P(\pi_{C_1}(x)) \frac{\prod_{k=2}^K P(\pi_{C_k}(x))}{\prod_{k=2}^K P(\pi_{S_k}(x))} \\ &= P(\pi_{C_1}(x)) \prod_{k=2}^K P(\pi_{C_k-S_k}(x) \mid \pi_{S_k}(x)) \end{aligned}$$

This is illustrated in the simple example of Figure 15.

Now if the graphical model is for the class conditional probability $P(x \mid \alpha)$, where α designates a class, x will be assigned to class γ where

$$\gamma = \operatorname*{argmax}_{\alpha} P(\pi_{C_1}(x) \mid \alpha) \prod_{k=2}^K P(\pi_{C_k-S_k}(x) \mid \pi_{S_k}(x), \alpha)$$

Notice that this is exactly of the form of a subspace ensemble classifier where the index sets $C_1,\ldots,C_K$ are the cliques of the conditional independence graph and designate the subspaces to which the measurement tuple x is projected. The method of combining is the product form which is equivalent to a sum form where log probabilities are used instead of probabilities.

How can this help in optimizing an ensemble classifier? Let $Q_1,\ldots,Q_M$ be the index sets designating the original subspaces of the ensemble of shallow classifiers. For each Q_m make a complete graph and union all the complete graphs together to form a graph G. From G, remove or add the smallest number of edges to make the modified G triangulated. Now find the cliques $C_1,\ldots,C_K$ of the triangulated graph G. Then, form K shallow classifiers based on projecting the measurement tuple to the subspaces designated by $C_1,\ldots,C_K$. When this is done, in the case of the N-tuple method, the N-tuple method becomes a graphical model for the class conditional probabilities. And since it is a graphical model, it carries with it the associated conditional independence assumptions. These conditional independence assumptions make it possible to understand what the combination of the marginal class conditional probabilities means. The combinations are just the class conditional probabilities for all the components of the measurement tuple.

Of course there is a difference between the ensemble classifiers and the graphical model classifiers. The idea of the ensemble classifier is to use many overlapping low dimensionality subspaces to define shallow and nearly independent classifiers. The idea of the graphical model classifiers is to define not quite as many overlapping low dimensionality subspaces to define the shallow subspace classifiers whose dependencies are taken into account by the inherent conditional independence assumptions. This is related to the version of the N-tuple classifier that uses conditional probabilities and whose combining method is by taking products of the class conditional probabilities or equivalently the sum of the logs of the class conditional probabilities. What this N-tuple classifier is doing is the computation of the products of the probabilities of the cliques. This is the numerator of (13). It misses the denominator of (13) which is the product of the separators.

9. Conclusion

For numerically valued variables, we have reviewed the properties of the correlation coefficient, the correlation ratio, the maximal correlation coefficient, and the monotone correlation coefficient. We gave examples illustrating some of the counterintuitive behavior of the maximal correlation coefficient and suggested that the monotone correlation coefficient may be more reasonable than the maximal correlation coefficient. We explored information theoretic measures of dependence, all of which are related to the mutual information between two variables. We discussed an adaptive partitioning method to estimate the mutual information in the case that the variables are continuously valued. We noted a few different ways that entropy and mutual information can be combined so that the result, a normalized form of mutual information is a metric. In addition, we listed a few such metrics which take values on the interval $[0,1]$.

We noted how the features of the cooccurrence probabilities included the maximal correlation coefficient and information theoretic measures of dependency. We discussed how cooccurrence probabilities can be used to define a joint probability image in which each pixel is the joint probability of the gray levels in the pixel's neighborhood. Neighborhoods can be regular such as 5×5 neighborhoods or can be entirely non-regular.

Finally, we discussed the constraint that defines a dependency. We took the case of the dependency between the feature vector and its ground truth class. Our setting was manifold methods. In particular,we discussed the N-tuple method and subspace classifiers and conjectured that there may be a universal approximation-like theorem by which the N-tuple and subspace classifiers might be shown to be able to approximate an arbitrary classification, under the constraint that the classification zones are sufficiently simple. And we have shown how the N-tuple classifier can be turned into a graphical model classifier.

Acknowledgments

Support under NASA grant NNX13AQ25G is gratefully acknowledged.

References

Aleksander, I., Morton, H., 1995. An Introduction to Neural Computing. International Thomson Computer Press, London.

Aleksander, I., T.J.Stonham, 1979. Guide to pattern recognition using random-access memories. IEE Journal on Computer and Digital Techniques 2, 29–40.

Allinson, N., Kolcz, A., 1997. N-tuple neural networks, in: Ellacott, S., Mason, J., Anderson, I. (Eds.), Mathematics of Neural Networks, pp. 3–14.

Baron, A., 1993. Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory 3, 930–945.

Bell, C., 1962. Mutual information and maximal correlation as measures of dependence. The Annals of Mathematical Statistics 33, 587–595.

Bledsoe, W., Browning, I., 1959. Pattern recognition and reading by machine, in: Proceeding Eastern Joint Computer Conference, Boston. pp. 232–255.

Brahma, P., Wu, D., She, Y., 2016. Why deep learning works: A manifold disentaglement perspective. IEEE Transactions On Neural networks and Learning Systems 27, 1997–2008.

Breiman, L., 1996. Bagging predictors. Machine Learning 24, 123–140.

Breiman, L., 2001. Random forests. Machine Learning 45, 5–32.

Breiman, L., Friedman, J., 1985. Estimating optimal transformations for multiple regression and correlation. Journal of the American Statistical Association 80, 580–598.

Chernyshov, K., 2015. Constructing consistent in the rényi sense measures of dependence within system identification, in: International Federation of Automatic Control, pp. 825–830.

Chow, C., Liu, C., 1968. Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory IT-14, 462–467.

Church, K., Hanks, P., 1990. Word association norms, mutual information and lexicography. Computational Linguistics , 22–29.

Cover, T., Thomas, J., 1991. Elements of Information Theory. John Wiley and Sons, Inc, New York.

Csàki, Fischer, J., 1963. On the general notion of maximum correlation. Magyar Tudomànyos Akad. Mat. Kutatò Intizètenk Közlemènyei 8, 27–51.

Cybenko, G., 1989. Approximation by superpositions of a sigmoidal function. Mathematics of Control Signals, and Systems 2, 303–314.

Darbellay, G., Vajda, I., 1999. Estimation of the information by an adaptive partitioning of the observation space. IEEE Transactions on Information Theory 45, 1315–1321.

Darbellay, G.A., Vajda, I., 2000. Entropy expressions for multivariate continuous distributions. IEEE Transactions on Information Theory 46, 709–712.

Dembo, A., Kagan, A., Shepp, L., 2001. Remarks on the maximum correlation coefficient. Bernoulli 7, 343–350.

Dietterich, T., 1998. An experimnetal comparison of threee methods of constructing ensembles of decision trees: Bagging, boosing, and randomization. Machine Learning , 1–22.

Dietterich, T., 2000. Ensemble methods in machine learning, in: Proceedings of the First International Workshop Multiple Classifier Systems, Springer-Verlag, Cagliari Italy. pp. 1–15.

Dobrushin, R., 1959. General formulation of shannon's main theorem in information theory. American Mathematical Socieituy Translations 33, 323–438.

Džeroski, S., Panov, P., Ženko, B., 2000. Machine learning, ensemble methods in, in: Encyclopedia of Complexity and System Science. Springer, pp. 5317–5325.

Etesami, O., Gohari, A., 2016. Maximal rank correlation. IEEE Communications Letters 20, 117–120.

Fedotov, A., Harremoes, P., Topsoe, F., 2003. Refinements of pinsker's inequality. IEEE Transactions on Infomration Theory 49, 1491–1498.

Feizi, S., Makhdoumi, A., Duffy, K., Kellis, M., Medard, M., 2015. Network maximal correlation. MIT-CSAIL-TR-2015-028 , 1–50.

Fowlkes, E., Kettenring, J., 1985. Estimating optimal transformation for multipel regression and correlation: Comment. Journal of the American Statistical Association 80, 607–613.

Freund, Y., Shapire, R., 1997. A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences 55, 119–139.

Gautam, A., Kimeldorf, G., 1999. Some results on the maximal correlation in $2 \times k$ contingency tables. The American Statistician 53, 336–341.

Ge, R., Zhou, M., Luo, Y., Meng, Q., Mai, G., Ma, D., Wang, G., Zhou, F., 2016. Mctwo: A two-step feature selection algorithm based on the maximal information coefficent. BMC Bioinformatics 17, 1–14.

Gebelein, H., 1941. Das statistische problem der korrelation als variations- und eigenwert-problem und sein zusammenhang mit der ausgleichungsrechnung. Zeitschrift für Angewwandte Mathematik und Mechanik 21, 364–379.

Ghazanfar, F., Ghani-Haider, N., 2016. Memory efficient sign language recognition system based on wisard weightless neural network technique, in: 2nd International Conference on Robotics and Artificial Intelligence, pp. 17–22.

Goodman, L., Kruskal, W., 1954. Measures of association for cross classification. Journal Of The American Statistical Association 49, 732–764.

Haralick, R., Diky, A., Su, X., Kiang, N., 2016. Inexact mdl for linear manifold clusters, in: 23rd International Conference on Pattern Recognition, pp. 1345–1351.

Haralick, R., Harpaz, R., 2005. Linear manifold clustering, in: International Workshop on Machine Learning and Data Mining in Pattern Recognition, Springer-Verlag. pp. 132–141.

Haralick, R., Harpaz, R., 2007. Linear manifold clustering in high dimensional spaces by stochastic search. Pattern Recognition 40, 2672–2684.

Haralick, R.M., 1975. A resolution preserving textural transform for image, in: Proceedings of the IEEE Computer Society Conference on Computer Graphics, Pattern Recognition and Data Structure, San Diego CA. p. 51=61.

Haralick, R.M., 1976. The table look up rule. Communications in Statistics, Theory, and Methods A5, 1163–1191.

Haralick, R.M., Shanmugam, K., Dinstein, I., 1973. Textural features for image classification. IEEE Transactions on Systems, Man, and Cybernetics SMC-3, 610–621.

Harpaz, R., Haralick, R., 2007. Linear manifold correlation clustering. International Journal of Information Technology And Intelligent Computing 2.

Hirschfeld, A., 1935. A connection between correlation and contingency. Proceedings of the Cambridge Philosophy Society 31, 520–524.

Ho, T.K., 1995. Random decision forests, in: Third International Conference on Document Analysis and Recognition, Montreal. pp. 14–18.

Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Transactions On Pattern Analysis and Machine Intelligence 20, 832–844.

Horibe, Y., 1985. Entropy and correlation. IEEE Transactions on Systems, Man, and Cybernetics SMC-15, 641–642.

Hornik, K., 1991. Approximation capabilities of multilayer feedfoward networks. Neural Networks 4, 251,257.

Hornik, K., Stinchcombe, M., White, H., 1989. Multi-layer feedforward networks are universal approximators. Neural Networks 2, 359–366.

Hsing, T., Liu, L.Y., Brun, M., Dougherty, E., 2005. The coefficient of intrinsic dependence (feature selection using el cid). Pattern Recognition 38, 623–636.

Jain, N., Murthy, C., 2016. A new estimate of mutual information based measure of dependence between two variables: Properties and fast implementation. International Journal of Machine Learning and Cybernetics 7, 857–875.

Jorgensen, T., Linneberg, C., 1999. Theoretical analysis and improved decision criteria for the n-tuple classifier. IEEE Transactions on Pattern Analysis and Machine Intelligence 21, 336–347.

Kabe, D., Gupta, A., 1990. On a multiple correlation ratio. Statistics and Probability Letters 9, 449–451.

Kim, H., Xu, J., Vemuri, B., Singh, V., 2015. Manifold-valued dirichlet processes, in: Proceedings of the 32nd International Conference on Machine Learning, Lille, France. pp. 1199–1208.

Kimeldorg, G., Sampson, A., 1978. Monotone dependence. The Annals of Statistics 6, 895–903.

Kinney, J., Atwal, G., 2014. Equitability, mutual information and the maximal information coefficient. Proceedings of the National Academy of Sciences 111, 3354–3359.

Kittler, J., Hatef, M., Duin, R., Matas, J., 1998. On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence 20, 226–239.

Kittler, J., Roli, F., 2000. Proceedings of the First International Workshop Multiple Classifier Systems. Springer, Cagliari Italy.

Kolcz, A., Allinson, N.M., 1996. N-tuple regression network. Neural Networks 9, 855–869.

Koller, D., Friedman, N., 2009. Probabilistic Graphical Models. MIT Press, Cambridge MA.

Kraskov, A., Stogbauer, H., Andrzejak, R., Grassberger, P., 2005. Hierarchical clustering using mutual information. Europhysics Letters 70, 278–284.

Kullback, S., 1959. Information Theory and Statistics. John Wiley and Sons, Inc., New York.

Kullback, S., Liebler, R., 1951. On information and sufficiency. Annals of Mathematical Statistics 22, 79–86.

Kumar, P., Hooda, D., 2008. On generalized measures of entropy and dependence. Mathematica Slovaca 58, 377–386.

Kvalseth, T., 1987. Entropy and correlation: Some comments. IEEE Transactions on Systems, Man, and Cybernetics SMC-17, 517–519.

Lancaster, H., 1957. Some properties of the bivariate normal distribution considered in the form of a contingency table. Biometrika 44, 289–292.

Lauritzen, S., 1996. Graphical Models. Claredon Press, Oxford.

Liang, Y., Shen, F., Zhao, J., Yang, Y., 2016. A fast manifold learning algortihms for dimensionality reductions, in: IEEE 28th International Conference on Tools with Artificial Intelligence, pp. 985–988.

Lin, J., 1991. Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory 37, 145–151.

Linfoot, E., 1957. An information measure of correlation. Information and Control 1, 85–89.

Liu, X., Srivastava, A., Gallivan, K., 2004. Optimal linear representations of images for object recognition. IEEE Transaction on Pattern Analysis and Machine Intelligence 26, 662–666.

Lucas, S., Amiri, A., 1995. Recognition of chain-coded handwwritten character images with scanning n-tuple method. Electronic Letters 31, 2088–2089.

Ludermir, T., Carvalho, A., Braga, A., Souto, M., 1999. Weightless neural models: A review of current and past works. Neural Computing Surveys 2.

Lui, Y.M., 2012. Advances in matrix manifolds for computer vision. Image and Vision Computing 30, 380–388.

Meilă, M., 2003. Comparing clusterings by the variation of information, in: Scholkopf, B., Warmuth, M. (Eds.), Learning Theory and Kernel Machines. Springer-Verlag, Berlin, pp. 173–187.

Meilă, M., 2005. Comparing clusterings - an axiomatic view, in: 22nd International Conference on Machine Learning, Bonn. pp. 577–584.

Meilă, M., 2007. Comparing clusterings – an information based distance. Multivariate Analysis 98, 873–895.

Meza, A.A., Lee, J., Verleysen, M., Castellanos-Dominguez, G., 2017. Kernel-based dimensionality reduction using renyi's α-entropy measures of similarity. Neurocomputing 222, 36–46.

Morciniec, M., Rohwer, R., 1995. The N-tuple Classifier: Too Good to Ignore. Technical Report. Aston University.

Nadarajah, S., Zografos, K., 2005. Expressions for rènyi and shannon entropies for bivariate distributions. Information Sciences 170, 173–189.

Nguyen, H., Muller, E., Vreeken, J., Efros, P., Bohm, K., 2014. Multivariate maximal correlation analysis, in: Proceedings of the 31^{st} International Conference on Machine Learning, pp. 775–783.

Nikitidis, S., Tefas, A., Pitas, I., 2014. Maximum margin projection subspace learning for visual data analysis. IEEE Transactions on Image Processing 23, 4413–4425.

Oja, E., 1983. Subspace Methods of Pattern Recognition. Research Studies Press, Letchworth UK.

Pearl, J., 1988. Probabilistic Reasoning in Intelligent Systems. Morgan Kaufmann, San Francisco.

Pinsker, M., 2005. On estimation of information via variation. Problems of Information Transmission 41, 71–75.

Prakash, M., Murty, N., 1996. Extended subspace methods of pattern recognition. Pattern Recognition Letters 17, 1131–1139.

Rényi, A., 1959. On measures of dependence. Acta Mathematica Science Hungarian 10, 218–226.

Rényi, A., 1961. On measures of entropy and information. Fourth Berkeley Symposium on Mathematical Statistics and Probability , 547–561.

Reshef, D., Reshef, Y., Finucane, H., Grossman, S., McVean, G., Turnbaugh, P., Lander, E., Mitzenmacher, M., Sabeti, P., 2011. Detecting novel association in large data sets. Science 334, 1518–1524.

Reza, F., 1961. An Introduction To Information Theory. McGraw-Hill, New York.

Rohwer, R., 1995. Two bayesian treatments of the n-tuple recognition method, in: Artificial Nerual Networks, pp. 171–176.

Rohwer, R., Morciniec, M., 1998. The theoretical and experimental status of the n-tuple classifier. Neural Networks 11, 1–14.

Rokach, L., 2010. Ensemble-based classifiers. Artificial Intelligence Review 33, 1–39.

Sampson, A., 1984. A multivariate correlation ratio. Statistics and Probability Letters 2, 77–81.

Srivastava, A., Klassen, E., 2004. Bayesian geometric subspace tracking. Journal For Advances In Applied Probability 36, 43–56.

Tattersall, G., Foster, S., Johnston, R., 1991. Single-layer lookup perceptrons. IEE Proceedings F-Radar and Signal Processing 138, 46–54.

Thanopoulos, A., fakotakis, N., Kokkinakis, G., 2002. Comparative evaluation of collocation extraction metrics, in: Language Resources Evaluation Conference, Las Palmas Spain. pp. 620–625.

Therrien, C., 1975. Eigenvalue properties of projection operators and their application to the subspace method of feature extraction. IEEE Transactions on Computers , 944–948.

Turaga, P., Veeraraghavan, A., Srivastava, A., Chellappa, R., 2011. Statistical computations on grassmann and stiefel manifolds for image and video based recognition. IEEE Transactions on Pattern Analysis and machine Intelligence 33, 2273–2286.

Valentini, G., Masulli, F., 2002. Ensembles of learning machines, in: 13th Italian Workshop on Neural Nets, Vietri sul Mare, Italy. pp. 3–22.

Vinh, N.X., Epps, J., Bailey, J., 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization, and correction for chance. Journal of Machine Learning Research 11, 2837–2854.

Watanabe, H., Katagiri, S., 1995. Discriminative subspace method for minimum error pattern recognition, in: IEEE Workshop on Neural Networks for Signal Processing, pp. 77–86.

Watanabe, S., 1960. Information theoretical analysis of multivariate correlation. IBM Journal of Research and Development 4, 67–82.

Watanabe, S., 1969. Knowing and Guessing. John Wiley and Sons, New York.

Watanabe, S., 1970. Feature compression, in: Tou, J. (Ed.), Advances in Information Sciences, Vol 3. Plenun Press, New York, pp. 63–111.

Watanabe, S., Pakvasa, M., 1973. Subspace method in pattern recognition, in: First International Joint Conference on Pattern Recognition, Washington DC. pp. 25–32.

Whittaker, J., 1990. Graphical Models in Applied Mutlvariate Statisticis. John Wiley and Sons, Inc., New York.

Witsenhausen, H., 1975. On sequences of pairs of dependent random variables. SIAM Journal on Applied Mathematics 28, 100–113.

Yin, M., Guo, Y., Gao, J., 2014. Linear subspace learning via sparse dimension reduction, in: International Joing Conference on Neural Networks, Beijing. pp. 3540–3547.

Yu, Y., 2008. On the maximal correlation coefficient. Statistics and Probability Letters 78, 1072–1075.

Yu, Z., Wang, D., You, J., Wong, H.S., Wu, S., Zhang, J., Han, G., 2016. Progressive subspace ensemble learning. Pattern Recognition 60, 692–705.

Zhang, Y., Jia, S., Huang, H., Qiu, J., Zhou, C., 2014. A novel algorithm for the precise calculation of the maximal information coefficient. Scientific Reports 4, 1–5.

Zhang, Z., 2007. Estimating mutual information via kolmogorov distance. IEEE Transactions On Information Theory 53, 3280–3282.