*Full paper*

# Symmetry Studies and Decompositions of Entropy

## Marlos Viana

The University of Illinois at Chicago Eye Center, 1855 W. Taylor St., Chicago, IL 60612, USA
E-mail: viana@uic.edu

**Abstract:** This paper describes a group-theoretic method for decomposing the entropy of a finite ensemble when symmetry considerations are of interest. The cases in which the elements in the ensemble are indexed by $\{1, 2, \ldots, n\}$ and by the permutations of a finite set are considered in detail and interpreted as particular cases of ensembles with elements indexed by a set subject to the actions of a finite group. Decompositions for the entropy in binary ensembles and in ensembles indexed by short DNA words are discussed. Graphical descriptions of the decompositions of the entropy in geological samples are illustrated. The decompositions derived in the present cases follow from a systematic data analytic tool to study entropy data in the presence of symmetry considerations.

## 1  Introduction

The entropy $H = -\sum_j p_j \log p_j$ of a finite set of n mutually exclusive events with corresponding probabilities $p_1, \ldots, p_n$ measures the amount of uncertainty associated with those events [1, p.3]. Its value is zero when any of the events is certain, it is positive otherwise, and attains its maximum value $(\log n)$ when the events are equally like, that is, $p_1 = \ldots = p_n = 1/n$. Alternatively [2, p.7], H is the mean value of the quantities $-\log p_j$ and can be interpreted as the mean information in an observation obtained to ascertain the mutually exclusive and exhaustive (hypotheses defined by those) events.

In general, the probabilities of an ensemble with n elements are indexed by the set

$$V = \{1, \ldots, n\},$$

and commonly indicated by $p_1, p_2, \ldots, p_n$. However, there are applications in which the elements in the ensemble exhibit intrinsic symmetries. Consider, for example, an election in which voters select their ordered preferences among the candidates in the list $\{a, b, c\}$ by choosing one of the 6 permutations in the set

$$V = \{abc, \quad cab, \quad bca, \quad bac, \quad cba, \quad acb\}.$$

The resulting distribution

$$p' = (p_{abc}, p_{cab}, \ldots, p_{acb})$$

of relative frequencies of those choices is then an example of empirical probabilities indexed by permutations. Accordingly, the corresponding entropy is written as

$$H = -\sum_{s \in V} p_s \log p_s.$$

In another example, the experimental results might be the empirical probabilities

$$p' = (p_{AA}, p_{AG}, \ldots, p_{TT})$$

with which the 16 binary sequences

$$V = \{AA, AG, \ldots, TT\}$$

in the DNA alphabet $\{A, G, C, T\}$ appear in a given reference sequence of interest. Similarly, then, these probabilities are indexed by a set of labels structurally more interesting than $\{1, 2, \ldots, n\}$. In fact, the 16 binary words can be classified, for example, by permutations shuffling the letters in the DNA alphabet, or distinctly, by permutations shuffling the positions of those letters. The objective of this paper is demonstrating that there is a systematic relationship among the labels, the (frequency) data indexed by those labels and the symmetries that are consistent with them, and showing that the methodology describing such relationship leads to the statistical analysis of the observed entropy of those frequency distributions.

The basic elements of symmetry studies are introduced in the next section and applied to the derivation of the standard (Section 3) and regular (Section 4) decompositions of entropy. In Section 5 a symmetry study of the entropy in Sloan Charts used in the quantification of visual acuity is presented. Additional comments are included in Section 6. The algebraic aspects present in this paper follow from the elements of the theory of linear representations of finite groups and can be found, for example, in [3, pp.1-24].

## 2   Symmetry Studies

The voting and the molecular biology data mentioned in the Introduction are examples of structured data in symmetry studies [4, 5]. These studies are centered on the notion of data (x) that are indexed by a finite set V of indices or labels (s) upon which certain symmetry transformations can be defined. Briefly, symmetry studies explore the symmetry transformations identified by the set of labels to facilitate the classification, interpretation and statistical analysis of the data $\{x(s), s \in V\}$ indexed by these labels. A finite group (G) with g elements acts on V and determines a linear representation ($\rho$) of G that operates in the data vector space ($\mathcal{V}$). The resulting factorization of $\mathcal{V}$ into a direct sum of invariant subspaces follows from the construction of algebraically orthogonal projections of the form $\mathcal{P} = n \sum_{\tau \in G} \chi(\tau^{-1}) \rho(\tau)/g$, one for each irreducible representation (in dimension of n and character $\chi$) of G. The canonical projections are the key elements leading to the explicit calculation and interpretation of the canonical invariants $\mathcal{P}_1 x, \mathcal{P}_2 x, \ldots$ in the data. If G has h irreducible representations, then there are h projections, and the identity operator I in $\mathcal{V}$ reduces according to the sum $I = \mathcal{P}_1 + \mathcal{P}_2 + \ldots + \mathcal{P}_h$ of algebraically orthogonal ($\mathcal{P}_i \mathcal{P}_j = \mathcal{P}_j \mathcal{P}_i = 0$ for $i \neq j$) projection ($\mathcal{P}_i^2 = \mathcal{P}_i$, $i = 1, \ldots, h$) matrices. A formal connection with the data analytical component of any symmetry study follows from the observation that basic decompositions of the form

$$(x|y) = (x|\mathcal{P}_1 y) + (x|\mathcal{P}_2 y) + \ldots + (x|\mathcal{P}_h y),$$

for any inner product $(\cdot|\cdot)$ defined in $\mathcal{V}$, can then always be studied within the context of statistical inference for quadratic forms [6, 7]. As a consequence, symmetry-related hypotheses defined by the canonical invariants can be identified and assessed [8–10].

The present paper concentrates on two basic types of canonical decompositions. In the *standard* case the components of the probability distributions $p' = (p_1, \ldots, p_n)$ are indexed by the set $V = \{1, \ldots, n\}$, the symmetry transformations are defined by the group $S_n$ of all permutations of those indices, which then acts on V according to

$$\tau : (1, 2, \ldots, n) \to (\tau 1, \tau 2, \ldots, \tau n), \quad \tau \in S_n.$$

In the *regular* case the components $(p_\sigma)$ of the probability distributions are indexed by the elements $\sigma$ of a finite group $(G, *)$ acting on itself according to

$$\tau : (\sigma)_{\sigma \in G} \to (\tau * \sigma)_{\sigma \in G}, \quad \tau \in G.$$

In closing this section it is opportune to remark that applications of group-theoretic principles in statistics and probability have a long history and tradition of their own, from Legendre-Gauss's least squares principle through R.A Fisher's method of variance decomposition. According to E.J. Hannan [11], the work [12] of A. James appears to be among the first describing the group-theoretic nature of Fisher's argument, giving meaning to the notion of relationship algebra of an experimental design. In that same early period, U. Grenander [13] showed the effectiveness of harmonic analysis techniques to extend classical limit theorems to algebraic structures such as locally compact groups, Banach spaces and topological algebras. Two decades later, the relevance of group invariance and group representation arguments in statistical inference would become evident [14–19]. The integral of Haar, object of L. Nachbin's 1965 monograph [20], became a familiar tool among statisticians. The work of S. Andersson on invariant normal models [21] is now recognized as a landmark concept extending and setting definitive boundaries to multivariate statistical analysis [22] in the tradition of T.W. Anderson. The collection of contemporary work in [23] clearly documents the present-day interest on those methods.

## 3   The Standard Decomposition of Entropy

Consider first a probability distribution p indexed by the set $V = \{1, 2\}$, that is $p' = (p_1, p_2)$. The group of permutations is $S_2 = \{1, t\}$, where 1 indicates the identity and t indicates the transposition (12). $S_2$ acts on V by permutation of the indices of their elements and the resulting linear representation

$$\tau : (1, 2) \xrightarrow{\rho(\tau)} (\tau 1, \tau 2), \quad \tau \in S_2,$$

is given by

$$\rho(1) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \rho(t) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

Note that $\rho(\tau)$ is the matrix defined by the change of basis $\{e_1, e_2\} \to \{e_{\tau 1}, e_{\tau 2}\}$ in the data vector space $\mathbb{R}^2$. The canonical projections associated with $\rho$ are defined by

$$\mathcal{A} = \frac{1}{2}[\rho(1) + \rho(t)] = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \mathcal{Q} = \frac{1}{2}[\rho(1) - \rho(t)] = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix},$$

where the coefficients $\{1, 1\}$ and $\{1, -1\}$ determining $\mathcal{A}$ and $\mathcal{Q}$ are the two irreducible characters of $S_2$. Note that

$$\mathcal{A}\mathcal{Q} = \mathcal{Q}\mathcal{A} = 0, \quad \mathcal{A}^2 = \mathcal{A}, \; \mathcal{Q}^2 = \mathcal{Q},$$

and

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \mathcal{A} + \mathcal{Q}.$$

Let $\ell' = (\log(p_1), \log(p_2))$, so that $H = -p'\ell$ is the entropy in the probability distribution p. It then follows from $I = \mathcal{A} + \mathcal{Q}$ that H decomposes as

$$H = -p'\ell = -p'I\ell = -(p'\mathcal{A}\ell + p'\mathcal{Q}\ell),$$

the components of which can be expressed as the log geometric mean

$$H_1 = -p'\mathcal{A}\ell = -\frac{1}{2}\log(p_1 p_2)$$

of the components of p, and as

$$H_2 = -p'\mathcal{Q}\ell = -\frac{1}{2}(p_1 - p_2)\log(\frac{p_1}{p_2}).$$

Because $H_2$ can be expressed as

$$H_2 = -\sum_{i=1}^{2}(p_i - \frac{1}{2})\log\frac{p_i}{1/2},$$

it then follows that $-H_2$ is precisely Kullback's [2, p.110] divergence between p and the uniform distribution $e'/2 = (1, 1)/2$, thus justifying the interpretation of entropy as a measure of non-uniformity.

The standard decomposition of the entropy obtained for two-component distributions extends to $n$−component distributions in the expected way. Writing

$$\mathcal{A} = \frac{1}{n}ee', \quad \mathcal{Q} = I - \mathcal{A},$$

where $ee'$ is the $n \times n$ matrix of ones, it then follows that the canonical decomposition is exactly

$$I = \mathcal{A} + \mathcal{Q}.$$

Clearly, it satisfies $\mathcal{A}^2 = \mathcal{A}$, $\mathcal{Q}^2 = \mathcal{Q}$ and $\mathcal{A}\mathcal{Q} = \mathcal{Q}\mathcal{A} = 0$. Moreover, $\mathcal{A}$ projects $\mathcal{V} = \mathbb{R}^n$ into a subspace $\mathcal{V}_a$ of dimension $\text{tr } \mathcal{A} = 1$ generated by $e = e_1 + \ldots + e_n = (1, 1, \ldots, 1)' \in \mathcal{V}$, whereas $\mathcal{Q}$ projects $\mathcal{V}$ into an irreducible subspace $\mathcal{V}_q$ in dimension $\text{tr } \mathcal{Q} = n - 1$, the orthogonal complement of $\mathcal{V}_a$ in $\mathcal{V}$. The irreducibility of $\mathcal{V}_q$, proved using character theory [3, p.17; 5, p.70], shows that the reduction $\mathcal{V} = \mathcal{V}_a + \mathcal{V}_q$ is exactly the canonical reduction determined by $\rho$. This decomposition is referred to as the *standard* decomposition or reduction. Applying the standard reduction $I = \mathcal{A} + \mathcal{Q}$ to the entropy $H = -p'\ell$ of a distribution $p' = (p_1, \ldots, p_n)$, where $\ell' = (\log(p1), \ldots, \log(p_n))$, it follows, in summary,

**Proposition 3.1** *The standard reduction of the entropy* H *in* $p' = (p_1, \ldots, p_n)$ *is* $H = H_1 + H_2$ *with*

$$H_1 = -\frac{1}{n}\log(p_1 p_2 \ldots p_n), \quad H_2 = -\frac{1}{n}\sum_{i<j}(p_i - p_j)\log\frac{p_i}{p_j}.$$

Similarly to the case $n = 2$ described earlier, $(-1\times)$ the entropy of a $n$−component distribution decomposes as the sum of the log geometric mean and Kullback's divergence

$$-H_2 = \sum_{i=1}^{n}(p_i - \frac{1}{n})\log\frac{p_i}{1/n}$$

between p and the uniform distribution $e'/n$. This can be easily verified by direct evaluation of the RHS of the above equality.

## 3.1   Graphical displays of $\{H_1, H_2\}$

It is observed that both $H_1$ and $H_2$ remain invariant under $S_n$, in the sense that for $\mathcal{P} = \mathcal{A}, \mathcal{Q}$,

$$(\rho(\tau)p)'\mathcal{P}(\rho(\tau)\ell) = p'\rho(\tau)'\mathcal{P}\rho(\tau)\ell = p'\mathcal{P}\ell. \qquad (3.1)$$

The last equality in (3.1) is a consequence of the fact that $\mathcal{P}$ and $\rho(\tau)$ commute for all $\tau \in S_n$. Therefore, $H_1$ and $H_2$ define a set of one-dimensional invariants that can be jointly displayed and interpreted, along with any additional covariates. Graphical displays such as these are generically called *invariant plots*.

Table (3.3) shows the observed frequencies with which the words in the permutation orbit

$$V = \{ACT, \ CTA, \ TAC, \ CAT, \ TCA, \ ATC\} \qquad (3.2)$$

of the DNA word ACT appear in 9 subsequent regions of the BRU isolate K02013 of the Human Immunodeficiency Virus Type I. To locate the sequence in the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov) data base, use the accession number K02013. Each region is in length of 900 base pairs. Table (3.3) also shows the entropy (H) of each distribution and its standard decomposition: the log geometric mean $(-H_1)$ and divergence $(-H_2)$ relative to the uniform distribution $e'/6$. In the present example $S_6$ acts on V by permutation of the six DNA words. In the next section the same frequency data will be studied under the permutation action of $S_3$ on the letters of the DNA words.
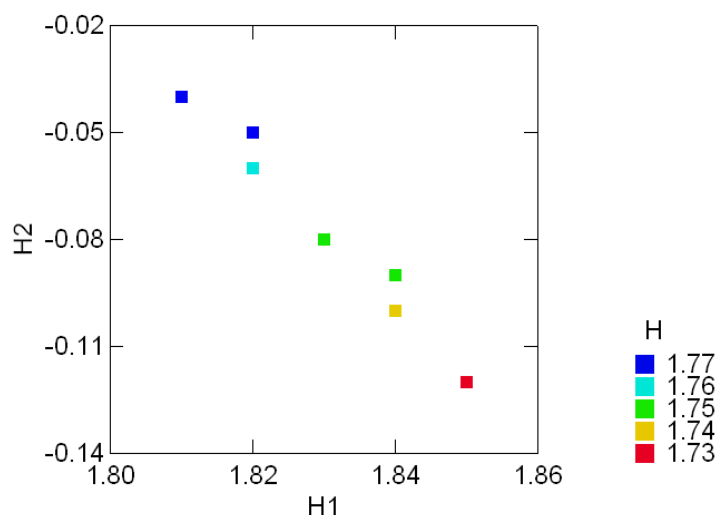
|       | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|-------|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| ACT | 8 | 16 | 16 | 7 | 17 | 11 | 12 | 6 | 14 |
| CTA | 15 | 8 | 14 | 9 | 14 | 15 | 8 | 5 | 16 |
| TAC | 7 | 17 | 13 | 15 | 9 | 11 | 18 | 5 | 17 |
| CAT | 14 | 15 | 16 | 14 | 21 | 17 | 15 | 10 | 8 |
| TCA | 11 | 18 | 10 | 17 | 11 | 16 | 14 | 9 | 13 |
| ATC | 7 | 15 | 9 | 13 | 11 | 11 | 11 | 12 | 10 |
| total | 62 | 89 | 78 | 75 | 83 | 81 | 78 | 47 | 78 |
| H | 1.74 | 1.76 | 1.77 | 1.75 | 1.75 | 1.77 | 1.76 | 1.73 | 1.76 |
| $H_1$ | 1.84 | 1.82 | 1.82 | 1.84 | 1.83 | 1.81 | 1.82 | 1.85 | 1.82 |
| $H_2$ | $-0.10$ | $-0.060$ | $-0.050$ | $-0.090$ | $-0.080$ | $-0.040$ | $-0.060$ | $-0.120$ | $-0.060$ |

(3.3)

The invariant plot of Figure 3.1 shows the (color coded) entropy of the ACT orbit and its invariant components $(H_1, H_2)$ in each one of nine regions. Note, for example, that the entropy in regions 3 and 6 are equal and yet their location in the invariant plot differ, showing a slightly increased divergence from the uniform distribution $e'/6$ in region 3. Although the entropy in regions 6 and 8 is essentially the same, there is a three-fold ratio between their divergence components- this is noticeable by inspecting the (lack of) uniformity in the corresponding frequency distributions.

Differentiations of that nature are possible because the dimension of the invariant subspace associated with $H_2$ is tr $\mathcal{Q} = n - 1$, where n is the number of components in the distribution under consideration. The regular decomposition described in the next section will further illustrate these differentiations and their relation to the symmetries imposed in the set of labels for the components of the distributions.

Figure 3.1: Color coded entropy levels (H) and their invariant components $\{H_1, H_2\}$ in 9 subsequent regions of the HIV virus Type I (BRU isolate).



### 3.2 The standard decomposition of the entropy of the Sloan fonts
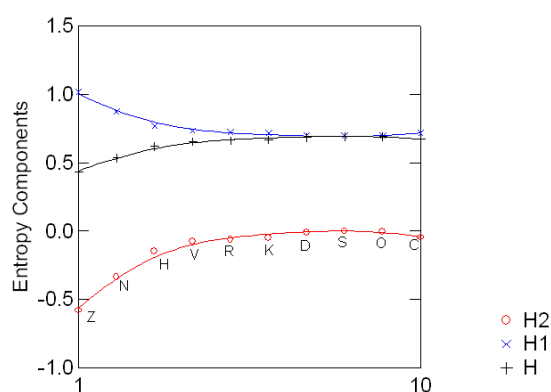
Table (3.4) shows the ten Sloan fonts [24, Table 5] used for the assessment of visual acuity in the Early Treatment Diabetic Retinopathy Study, along with their estimated difficulty (probability of incorrectly identifying the letter), the corresponding entropy and their standard invariant components $H_1$ and $H_2$. Figure 3.2 shows the standard decomposition for each letter. The divergence from the uniform distribution e'/2 ranges from 0.58 when the distribution is $(0.844, 0.156)$ to 0.001

for $(0.516, 0.484)$. The study of the entropy of Sloan fonts will continue later on in Section 5.

| Sloan Font | Difficulty | Entropy | $H_1$ | $H_2$ |
|:----------:|:----------:|:-------:|:-----:|:-----:|
| Z | 0.844 | 0.433 | 1.01 | $-0.580$ |
| N | 0.774 | 0.535 | 0.870 | $-0.337$ |
| H | 0.688 | 0.619 | 0.770 | $-0.152$ |
| V | 0.636 | 0.656 | 0.730 | $-0.076$ |
| R | 0.622 | 0.663 | 0.725 | $-0.061$ |
| K | 0.609 | 0.669 | 0.720 | $-0.048$ |
| D | 0.556 | 0.687 | 0.700 | $-0.012$ |
| S | 0.516 | 0.693 | 0.695 | $-0.001$ |
| O | 0.47 | 0.692 | 0.695 | $-0.003$ |
| C | 0.393 | 0.673 | 0.715 | $-0.040$ |

$$(3.4)$$

Figure 3.2: Entropy canonical components for the ten Sloan fonts.
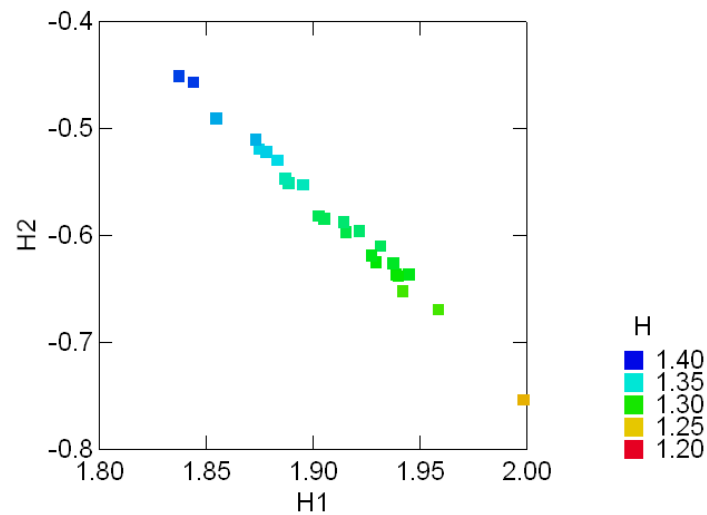


### 3.3 Geological compositions

Tables .1 in Appendix A appear in [25, pp. 354, 358] and show the geological compositions of albite, blandite, cornite, daubite and endite in 25 samples of coxite and hongite, in addition to the porosity (percentage of void space) of each sample of coxite. In this example, probability distributions are referred to as *compositions*, following [25, p. 26]. Figures 3.3 and 3.4 show, respectively, the color coded entropy and porosity levels for the coxite samples, and their invariant

components $\{H_1, H_2\}$. The entropy among all compositions is within the range of $75 - 87$ percent of the max entropy ($\log 5 = 1.6$), thus being concentrated in a relatively narrow region. The joint distribution of porosity and $H_2$ shown in Figure 3.5 suggests that porosity is negatively correlated with $H_2$, or, equivalently, positively correlated with the divergence. In fact, the observed sample correlation coefficient based on 25 samples of these two variables is 0.78.

Figure 3.6 shows the color coded entropy in the hongite samples and their invariant components $\{H_1, H_2\}$. Figure 3.7 clearly shows the contrasting difference in entropy range between the two minerals, also evident in the range of the divergence in those compositions. This noticeable feature of the compositions[25, p.4], namely that coxite compositions are much less variable than hongite compositions, is clearly captured in these invariant plots.
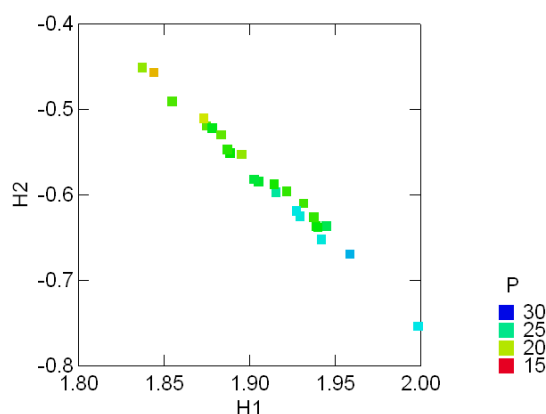
Figure 3.3: Color coded entropy levels (H) and their invariant $\{H_1, H_2\}$ components in the geological compositions of 25 samples of coxite.



# 4   Regular Decompositions of Entropy

In this section, the decomposition of the entropy for probability ensembles with components indexed by $S_3$, the group of permutations of three objects, will be considered. Decompositions for ensembles indexed by other finite groups can be obtained similarly.

Figure 3.4: Color coded porosity levels (P) and their invariant entropy components $\{H_1, H_2\}$ in the geological compositions of 25 samples of coxite.



To illustrate, consider the DNA word ACT and its permutation orbit

$$V = \{ACT,\ CTA,\ TAC,\ CAT,\ TCA,\ ATC\}$$

introduced early on in Section 3.1. Note that V can be generated from the word ACT by applying, respectively, the permutations

$$\begin{pmatrix} A\ C\ T \\ A\ C\ T \end{pmatrix},\ \begin{pmatrix} A\ C\ T \\ C\ T\ A \end{pmatrix},\ \begin{pmatrix} A\ C\ T \\ T\ A\ C \end{pmatrix},\ \begin{pmatrix} A\ C\ T \\ C\ A\ T \end{pmatrix},\ \begin{pmatrix} A\ C\ T \\ T\ C\ A \end{pmatrix},\ \begin{pmatrix} A\ C\ T \\ A\ T\ C \end{pmatrix}$$

defining the group $S_3$, so that V is isomorphic to $S_3$. Indicate by

$$p' = (r_1, r_2, r_3, t_1, t_2, t_3)$$

the respective relative frequencies with which the words in V, or equivalently, the elements in $S_3$, appear in a given DNA reference region. These relative frequencies are, consequently, data indexed by $S_3$. Let also $\ell$ indicate the vector of the log components of p, so that $H = -p'\ell$ is the entropy in the probability distribution p indexed by $S_3$.

## 4.1 The regular decomposition
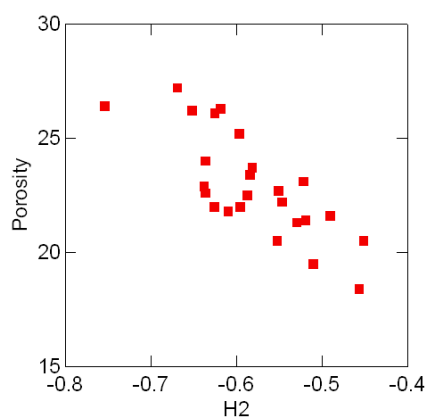
The permutations in $S_3$ act on the DNA words by shuffling the letters. This gives a linear representation

$$\tau : (\sigma)_{\sigma \in G} \xrightarrow{\rho(\tau)} (\tau * \sigma)_{\sigma \in G}, \quad \tau \in G = S_3$$

$\rho$ of $S_3$. Each matrix $\rho(\tau)$ is defined in the vector space for the data indexed by $S_3$. Associated with $\rho$ there are three canonical projections

$$\mathcal{P}_i = \frac{n_i}{g} \sum_{\tau \in G} \chi_i(\tau^{-1}) \rho(\tau), \quad i = 1, 2, 3,$$

Figure 3.5: Porosity levels and divergence components $H_2$ in 25 samples of coxite.



corresponding to the irreducible characters

$$
\left[
\begin{array}{c|ccc}
\chi & 1 & t & r \\
\hline
\chi_1 & 1 & 1 & 1 \\
\chi_2 & 2 & 0 & -1 \\
\hline
\chi_3 & 1 & -1 & 1
\end{array}
\right],
$$
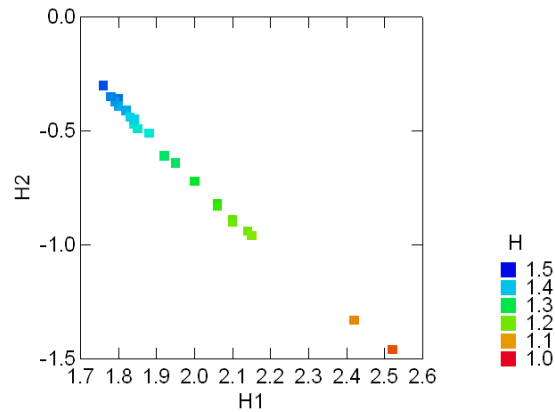
of $S_3$, with $n_1 = n_3 = 1$, $n_2 = 2$, $g = 6$, $t = \{(12),(13),(23)\}$, $r = \{(123),(132)\}$, given by

$$
\mathcal{P}_1 = 1/6
\begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1
\end{bmatrix},
\quad
\mathcal{P}_2 = 1/3
\begin{bmatrix}
2 & -1 & -1 & 0 & 0 & 0 \\
-1 & 2 & -1 & 0 & 0 & 0 \\
-1 & -1 & 2 & 0 & 0 & 0 \\
0 & 0 & 0 & 2 & -1 & -1 \\
0 & 0 & 0 & -1 & 2 & -1 \\
0 & 0 & 0 & -1 & -1 & 2
\end{bmatrix},
$$

and

$$
\mathcal{P}_3 = 1/6
\begin{bmatrix}
1 & 1 & 1 & -1 & -1 & -1 \\
1 & 1 & 1 & -1 & -1 & -1 \\
1 & 1 & 1 & -1 & -1 & -1 \\
-1 & -1 & -1 & 1 & 1 & 1 \\
-1 & -1 & -1 & 1 & 1 & 1 \\
-1 & -1 & -1 & 1 & 1 & 1
\end{bmatrix}.
$$

Figure 3.6: Color coded entropy levels (H) and their invariant $\{H_1, H_2\}$ components in the geological compositions of 25 samples of hongite.



Observe that $\mathcal{P}_i\mathcal{P}_j = \mathcal{P}_j\mathcal{P}_i = 0$ for $i \neq j$, $\mathcal{P}_i^2 = \mathcal{P}_i$, $i = 1, 2, 3$ and $I = \mathcal{P}_1 + \mathcal{P}_2 + \mathcal{P}_3$. The underlying invariant image subspaces $\mathcal{P}_i x$, $x \in \mathbb{R}^6$, are in dimension of $1, 4, 1$ respectively.

It then follows that

$$H = -p'\ell = -p'\mathcal{P}_1\ell - p'\mathcal{P}_2\ell - p'\mathcal{P}_3\ell,$$

with the regular components of the entropy H given by

$$H_1 = -p'\mathcal{P}_1\ell = -\frac{1}{6}\log[r_1 r_2 r_3 t_1 t_2 t_3],$$

$$H_2 = -p'\mathcal{P}_2\ell = -\frac{1}{3}\log[(\frac{r_1}{r_2})^{r_1-r_2}(\frac{r_1}{r_3})^{r_1-r_3}(\frac{r_2}{r_3})^{r_2-r_3}(\frac{t_1}{t_2})^{t_1-t_2}(\frac{t_1}{t_3})^{t_1-t_3}(\frac{t_2}{t_3})^{t_2-t_3}],$$

and

$$H_3 = -p'\mathcal{P}_3\ell = -\frac{1}{6}\log(\frac{r_1 r_2 r_3}{t_1 t_2 t_3})^{(r_1+r_2+r_3-t_1-t_2-t_3)}.$$
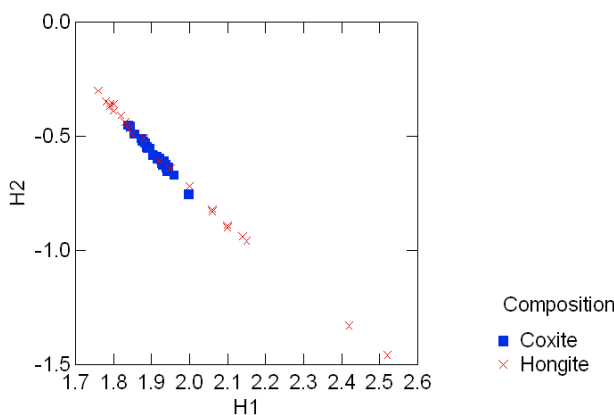
The interpretation of these components can be expressed in terms of three axial reflections (transpositions) and the three-fold rotations (cyclic permutations) of the regular triangle: $-H_1$ is the log geometric mean of the components of p; direct evaluation shows that

$$-H_2 = r_\bullet D(r : u) + t_\bullet D(t : u)$$

where $r_\bullet$ is the marginal probability $r_1 + r_2 + r_3$ of a rotation, $t_\bullet$ is the marginal probability $t_1 + t_2 + t_3$ of a reflection, $D(r : u)$ is Kullback's divergence between the rotation subcomposition [25, p.33]

$$r = (r_1, r_2, r_3)/(r_1 + r_2 + r_3)$$

Figure 3.7: Entropy levels in the standard invariant plot for coxite and hongite data.



and the uniform distribution $u = e'/3$, and $D(t : u)$ is the divergence between the reflection subcomposition $t = (t_1, t_2, t_3)/(t_1 + t_2 + t_3)$ and the uniform distribution $u$; and

$$-H_3 = \frac{1}{3}D((r, t) : (1/2, 1/2)) + \frac{1}{2}(r_\bullet - t_\bullet)[\log \text{geom. mean } (r) - \log \text{geom. mean } (t)].$$

It is then possible to summarize the regular decomposition of the entropy of a distribution $p$ indexed by $S_3$ as a three-component sum: the log geometric mean of $p$, the total uniform divergence within rotations and within reflections and a component measuring the between rotations and reflections separation.

Table (4.1) shows the regular decomposition $H = H_1 + H_2 + H_3$ for the permutation orbit of ACT in each one of 9 adjacent regions of the isolate. Table (4.2) shows the ratios $H_1/H$ and $H_2/H_3$ for potential comparisons among these regions. Similar ratios were calculated for the words AGT and ACG (Tables (4.3) and (4.4), respectively). The regions 1 and 4 were removed from the calculations for ACG due to presence of zeros in the frequency distributions.

|       | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|-------|---|---|---|---|---|---|---|---|---|
| $H$   | 1.74 | 1.76 | 1.77 | 1.75 | 1.75 | 1.77 | 1.76 | 1.73 | 1.76 |
| $H_1$ | 1.84 | 1.82 | 1.82 | 1.84 | 1.83 | 1.81 | 1.82 | 1.85 | 1.82 |
| $H_2$ | $-0.0969$ | $-0.0505$ | $-0.0338$ | $-0.0506$ | $-0.0836$ | $-0.0289$ | $-0.0607$ | $-0.0120$ | $-0.0195$ |
| $H_3$ | $-0.0013$ | $-0.0081$ | $-0.0120$ | $-0.0343$ | $-0.0010$ | $-0.0072$ | $-0.0012$ | $-0.1050$ | $-0.0444$ |

(4.1)

| ACT | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|-----|---|---|---|---|---|---|---|---|---|
| $H_1/H$ | 1.06 | 1.03 | 1.03 | 1.05 | 1.05 | 1.02 | 1.03 | 1.07 | 1.03 |
| $H_2/H_3$ | 72.3 | 6.20 | 2.82 | 1.48 | 81.2 | 4.01 | 49.0 | 0.114 | 0.439 |

(4.2)

| AGT | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $H_1/H$ | 1.03 | 1.02 | 1.03 | 1.10 | 1.06 | 1.05 | 1.09 | 1.03 | 1.01 |
| $H_2/H_3$ | 2.76 | 2.70 | 7.43 | 0.222 | 108.0 | 1.72 | 1.56 | 2.07 | 4.03 |

(4.3)

| ACG | 2 | 3 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|
| $H_1/H$ | 1.58 | 1.59 | 1.50 | 1.56 | 1.44 | 1.17 | 1.24 |
| $H_2/H_3$ | 1.45 | 0.614 | 0.736 | 0.295 | 0.404 | 0.313 | 0.324 |

(4.4)

The within-region and between-words in the ratios of the entropy regular components for the three DNA words considered in this example is striking. In particular, the ratio $H_2/H_3$ clearly differentiates the ACG orbit from ACT and AGT. The between-region range of the ratios $H_2/H_3$ is $0.295 - 1.45$ in the ACG orbits, in contrast to $0.222 - 108$ in the AGT orbits and $0.114 - 72.3$ in the ACT orbits.

## 5   A Symmetry Study of Sloan Charts

This study will combine the notions developed in the previous sections by introducing the Sloan Charts and considering each line in the chart as the sampling unit (total of 42 lines); identifying the group G of planar symmetries of the Sloan fonts; determining the invariants associated with the regular representation of G and studying the line entropy in that space. Cyclic permutation orbits have been used [26] to describe the evolutionary strategy of the HIV-1 virus.

### 5.1   The Sloan Charts and the Sloan Fonts symmetries

Tables (5.2) show a Sloan Chart [24, Table 5] developed for use in the Early Treatment Diabetic Retinopathy Study and the 10 individual Sloan letters appearing in Example 3.2, their symmetry transformations, the single-letter difficulty as the estimated [24] probability of incorrectly identifying the letter, and the corresponding entropy.

The group of interest here is the point planar group $G = \{1, o, v, h\}$, in which 1 is the identity, o the point symmetry or inversion, v the vertical axis reflection and h the horizontal axis reflection. The font is considered *centered* at the center of inversion with their natural vertical and horizontal directions along the corresponding axes of reflection. Observe that the symmetries indicated in the RHS Table in (5.2) are the subgroups of G leaving the letter fixed (or the letter stabilizer). G is an Abelian group isomorphic to $C_2 \times C_2$ and its multiplication table is given by

| $*$ | 1 | v | h | o |
|---|---|---|---|---|
| 1 | 1 | v | h | o |
| v | v | 1 | o | h |
| h | h | o | 1 | v |
| o | o | h | v | 1 |

(5.1)

$$
\begin{array}{ccccc}
C & O & H & Z & V \\
S & Z & N & D & C \\
V & K & C & N & R \\
K & C & R & H & N \\
Z & K & D & V & C \\
H & V & O & R & K \\
R & H & S & O & N \\
K & S & V & R & H \\
H & N & K & C & D \\
N & D & V & K & O \\
D & H & O & S & Z \\
V & R & N & D & O \\
C & Z & H & K & S \\
O & R & Z & S & K
\end{array}
$$

| Letter | Symmetries | Difficulty | Entropy |
|:------:|:----------:|:----------:|:-------:|
| Z | $1, o$ | 0.844 | 0.433 |
| N | $1, o$ | 0.774 | 0.535 |
| H | $1, o, v, h$ | 0.688 | 0.619 |
| V | $1, v$ | 0.636 | 0.656 |
| R | $1$ | 0.622 | 0.663 |
| K | $1$ | 0.609 | 0.669 |
| D | $1, h$ | 0.556 | 0.687 |
| S | $1, o$ | 0.516 | 0.693 |
| O | $1, o, v, h$ | 0.470 | 0.692 |
| C | $1, h$ | 0.393 | 0.673 |

$$(5.2)$$

The objective here is studying the relationship between line entropy and letter symmetry based on 42 ETDRS lines, obtained from three charts similar to the one shown in the LHS table in (5.2).

## 5.2 The regular decomposition

The canonical projections for the regular representation (G acting on itself) are given by

$$
\mathcal{P}_1 = 1/4 \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad
\mathcal{P}_o = 1/4 \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix},
$$

$$
\mathcal{P}_v = 1/4 \begin{bmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{bmatrix}, \quad
\mathcal{P}_h = 1/4 \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.
$$

Indicating, for simplicity of notation, by $x' = (1, o, v, h)$ a generic vector of data indexed by G, direct evaluation shows that

$$
\mathcal{P}_1 x = \frac{1}{4}\hat{x}_1[1, 1, 1, 1]', \quad \mathcal{P}_o x = \frac{1}{4}\hat{x}_o[1, 1, -1, -1]',
$$

$$\mathcal{P}_{\mathrm{v}}\mathrm{x} = \frac{1}{4}\widehat{\mathrm{x}}_{\mathrm{v}}[1, -1, 1, -1]', \quad \mathcal{P}_{\mathrm{h}}\mathrm{x} = \frac{1}{4}\widehat{\mathrm{x}}_{\mathrm{h}}[1, -1, -1, 1]',$$

where $\widehat{\mathrm{x}}$ are the Fourier transforms

$$\widehat{\mathrm{x}}_1 = 1 + \mathrm{o} + \mathrm{v} + \mathrm{h}, \quad \widehat{\mathrm{x}}_{\mathrm{o}} = 1 + \mathrm{o} - \mathrm{v} - \mathrm{h},$$

$$\widehat{\mathrm{x}}_{\mathrm{v}} = 1 + \mathrm{v} - \mathrm{o} - \mathrm{h}, \quad \widehat{\mathrm{x}}_{\mathrm{h}} = 1 + \mathrm{h} - \mathrm{o} - \mathrm{v},$$

evaluated at the irreducible one-dimensional representations of G.

Note, as shown in Appendix B, that

$$(\mathcal{P}_{\beta}\mathrm{x})_{\sigma} = \frac{\mathrm{n}_{\beta}}{\mathrm{g}}\mathrm{tr}\ [\beta(\sigma^{-1})\widehat{\mathrm{x}}(\beta)] = \mathrm{x}_{\beta}(\sigma),$$

that is, the entry $\mathrm{x}(\tau)$ of the regular $\beta$-invariant $\mathcal{P}_{\beta}\mathrm{x}$ is precisely $\mathrm{x}_{\beta}(\tau) = \mathrm{n}_{\beta}\mathrm{tr}\ [\beta(\tau^{-1})\widehat{\mathrm{x}}(\beta)]/\mathrm{g}$. Consequently, the data assignment $\mathrm{x}_{\beta}$ identifies the properties of the regular canonical projection $\mathcal{P}_{\beta}$ with the properties of the Fourier transform $\widehat{\mathrm{x}}(\beta)$ evaluated at $\beta$. In that sense, the indexing $\mathrm{x}_{\beta}$ should retain the interpretations associated with the invariant subspace of $\mathcal{P}_{\beta}$, and, jointly, the data vectors $\{\mathrm{x}_{\beta}, \beta \in \widehat{\mathrm{G}}\}$ should fully describe the regular symmetry invariants. Here $\widehat{\mathrm{G}}$ indicates the set of all irreducible representations of G. **Any** indexing x of G decomposes (via the Fourier inverse formula) as the linear superposition $\sum_{\beta \in \widehat{\mathrm{G}}} \mathrm{x}_{\beta}(\tau)$ of regular invariants $\mathrm{x}_{\beta}$ of G. Shortly, then $\mathrm{x} = \sum_{\beta \in \widehat{\mathrm{G}}} \mathrm{x}_{\beta}$.

*5.3   Sorting the line entropy by font symmetry type*

Indicate by
$$\mathrm{fix}\ (\tau) = \{\mathrm{s} \in \mathrm{V}; \varphi(\tau, \mathrm{s}) = \mathrm{s}\} \subseteq \mathrm{V}$$

the set of elements in V that remain fixed by the symmetry transformation $\tau$ applied to $\mathrm{s} \in \mathrm{V}$ according to the rule $\varphi$. Then there is a general method of indexing the data by G, called the regular indexing, constructed as follows: to each element $\tau \in \mathrm{G}$ associate the evaluation $\underline{\mathrm{x}}(\tau)$ of a scalar summary of x defined over the set fix $(\tau)$. That is, $\underline{\mathrm{x}}(\tau)$ indicates a summary of the data over those elements (if any) in V that share the symmetry of $\tau$. For example, if the summary of interest is the averaging, then

$$\underline{\mathrm{x}}(\tau) = \frac{1}{|\mathrm{fix}(\tau)|} \sum_{\mathrm{s} \in \mathrm{fix}(\tau)} \mathrm{x}(\mathrm{s}), \quad |\mathrm{fix}(\tau)| > 0.$$

As an example, the regular indexing is applied to the font entropy data $\mathrm{x}(\mathrm{s})$ of each font s in each line (V) of the Sloan Chart shown in the LHS Table in (5.2). Here $\mathrm{fix}(\tau)$ is the set of fonts with the symmetry of $\tau \in \mathrm{G}$, and $\underline{\mathrm{x}}(\tau)$ is the average font entropy over the set of fonts with the symmetry of $\tau$, or the mean line entropy conditional to fonts with symmetry of $\tau$. Table (5.3) shows the line entropy indexed, or sorted, by the symmetries of G. Note that line 2 was excluded
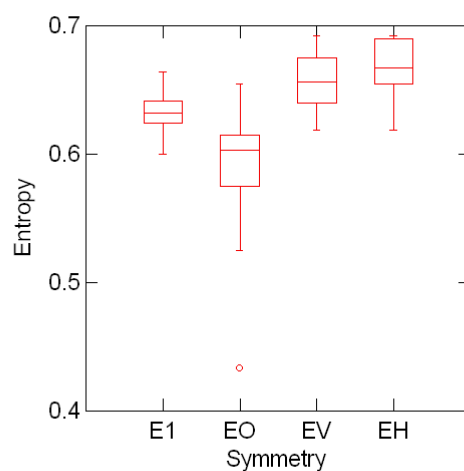
because there were vertical symmetries in that line. One can then proceed with the study of the regular decomposition for $\underline{x}$ as data indexed by the group G of planar symmetries.

| Sloan chart line | $\underline{x}(1)$ | $\underline{x}(o)$ | $\underline{x}(v)$ | $\underline{x}(h)$ |
|:---:|:---:|:---:|:---:|:---:|
| COHZV | 0.614 | 0.580 | 0.657 | 0.660 |
| VKCNR | 0.636 | 0.535 | 0.656 | 0.673 |
| KCRHN | 0.632 | 0.575 | 0.619 | 0.645 |
| ZKDVC | 0.624 | 0.433 | 0.656 | 0.680 |
| HVORK | 0.660 | 0.655 | 0.657 | 0.655 |
| RHSON | 0.640 | 0.635 | 0.655 | 0.655 |
| KSVRH | 0.660 | 0.655 | 0.640 | 0.619 |
| HNKCD | 0.636 | 0.575 | 0.619 | 0.660 |
| NDVKO | 0.648 | 0.615 | 0.675 | 0.690 |
| DHOSZ | 0.624 | 0.608 | 0.655 | 0.667 |
| VRNDO | 0.648 | 0.615 | 0.675 | 0.690 |
| CZHKS | 0.618 | 0.580 | 0.619 | 0.645 |
| ORZSK | 0.630 | 0.603 | 0.692 | 0.692 |

(5.3)
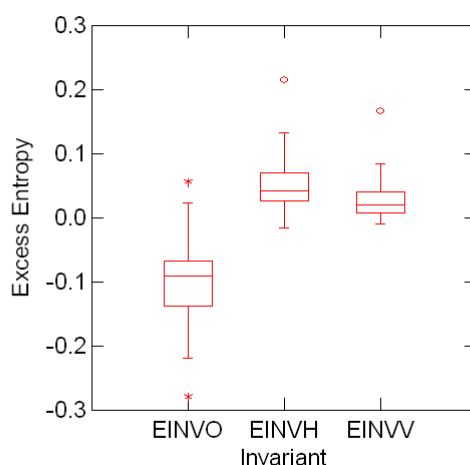
Figure 5.1 shows the line entropy distribution classified by font symmetry $E_\tau \equiv \underline{x}(\tau)$ over 39 Sloan

Figure 5.1: Mean line entropy distribution and font symmetry $E_\tau \equiv \underline{x}(\tau)$.



charts lines. In three of the 42 lines the frequency counts were not all positive and those lines were

Figure 5.2: Distribution of the entropy invariants EINVO$\equiv \widehat{x}_o$, EINVV$\equiv \widehat{x}_v$ and EINVH$\equiv \widehat{x}_h$ for the entropy data indexed by font symmetries.



deleted. The results suggest a (statistically) significant drop in mean line entropy conditional to fonts with point symmetry. This result should be explored further taking into account, for example, the fact that the mean number of letters per line ($\pm$ standard deviation) with point, horizontal and vertical symmetry are, respectively, $2.5 \pm 0.804$, $2.0 \pm 0.698$, and $1.5 \pm 0.741$.

In addition, Figure 5.2 shows the distribution of the Fourier transforms $\widehat{x}_2$, $\widehat{x}_3$ and $\widehat{x}_4$ (second, third and fourth invariants) for the entropy data

$$\{\underline{x}(1), \underline{x}(o), \underline{x}(v), \underline{x}(h)\}$$

of Table (5.3) indexed by the font symmetries. Consistently with the previous interpretation, the distribution of the invariant $\widehat{x}_o$ is markedly shifted from the distributions of $\widehat{x}_v$ and $\widehat{x}_h$.

## 6  Summary and Comments

This paper introduced the *standard* decomposition H $=$ H$_1$ + H$_2$ of the entropy of any finite distribution and the method for obtaining the *regular* decomposition H $=$ H$_1$ + ... for the entropy of distributions indexed by arbitrary finite groups.

1. The standard decomposition appears in analogy with the decomposition of the sum of squares, or, more specifically, to the decomposition of an inner product (y|x) of two vector x and y in the same finite-dimensional vector space. In fact, the standard decomposition $I = \mathcal{A} + \mathcal{Q}$ has a well-known role in statistics. It leads to the usual decomposition

$$x'x = \sum x_i^2 = n(\overline{x})^2 + \sum_i (x_i - \overline{x})^2$$

of the sum of squares in terms of the sample average $\overline{x}$ and the sample variance $\frac{1}{n}\sum_i (x_i - \overline{x})^2$. Similarly, it decomposes $x'y$ into the sum $n\overline{x}\,\overline{y} + \sum(x_i - \overline{x})(y_i - \overline{y})$;

2. The methodology presented in this paper should provide an additional tool to study the entropy in distributions of nucleotide sequences in molecular biology data. This case is also of statistical and algebraic interest because it extends the decompositions introduced in the present paper to the case in which the probability distributions are indexed by the short nucleotide sequences upon which a group may act by symbol symmetry or by position symmetry [5, p.40];

3. Two-dimensional invariant plots for the regular decomposition of the entropy by $S_3$ can be obtained by jointly displaying the three pairwise combinations of the invariant components $\{H_1, H_2, H_3\}$. The characteristic of the regions of constant entropy in these planes needs to be investigated;

4. The large-sample theory for multinomial distributions, described in detail, for example, in [27, p.469], can be applied to derive the moments of the entropy components.

# Appendix A

**Coxite Data**

| albite | blandite | cornite | daubite | endite | porosity |
|---|---|---|---|---|---|
| 44.2 | 31.9 | 5.40 | 10.5 | 8.0 | 21.8 |
| 4.90 | 25.4 | 5.80 | 11.3 | 8.50 | 25.2 |
| 50.2 | 24.8 | 5.70 | 11.1 | 8.20 | 26.1 |
| 49.9 | 24.7 | 5.40 | 11.4 | 8.60 | 26.3 |
| 48.5 | 27.8 | 5.90 | 10.2 | 7.60 | 22.6 |
| 45.9 | 27.1 | 6.90 | 11.5 | 8.60 | 21.4 |
| 44.1 | 31.9 | 6.0 | 10.2 | 7.80 | 22.0 |
| 46.4 | 29.9 | 5.50 | 10.3 | 7.90 | 22.0 |
| 45.7 | 27.0 | 6.20 | 12.0 | 9.10 | 23.1 |
| 46.4 | 30.0 | 5.10 | 10.4 | 8.10 | 24.0 |
| 41.7 | 30.2 | 7.70 | 11.6 | 8.80 | 18.4 |
| 44.9 | 25.7 | 7.70 | 12.4 | 9.30 | 20.5 |
| 48.6 | 27.7 | 5.80 | 10.2 | 7.70 | 22.9 |
| 49.7 | 26.7 | 4.90 | 10.6 | 8.10 | 27.2 |
| 49.6 | 24.4 | 6.40 | 11.2 | 8.40 | 23.4 |
| 46.5 | 28.6 | 5.90 | 10.7 | 8.30 | 22.5 |
| 47.3 | 24.2 | 7.90 | 11.8 | 8.80 | 21.6 |
| 44.7 | 30.0 | 6.80 | 10.5 | 8.0 | 20.5 |
| 48.0 | 25.6 | 7.0 | 11.1 | 8.30 | 22.7 |
| 50.0 | 23.8 | 6.60 | 11.2 | 8.40 | 23.7 |
| 51.4 | 24.2 | 5.70 | 10.7 | 8.0 | 26.2 |
| 53.3 | 25.1 | 5.20 | 9.40 | 7.0 | 26.4 |
| 47.9 | 25.4 | 6.70 | 11.4 | 8.60 | 22.2 |
| 43.5 | 29.8 | 6.70 | 11.2 | 8.80 | 19.5 |
| 44.5 | 29.2 | 6.50 | 11.2 | 8.60 | 21.3 |

**Hongite Data**

| albite | blandite | cornite | daubite | endite | H | $H_1$ | $H_2$ |
|---|---|---|---|---|---|---|---|
| 0.488 | 0.317 | 0.0380 | 0.0640 | 0.0930 | 1.24 | 2.06 | −0.82 |
| 0.482 | 0.238 | 0.0900 | 0.0920 | 0.0980 | 1.36 | 1.85 | −0.49 |
| 0.370 | 0.0910 | 0.342 | 0.0950 | 0.102 | 1.41 | 1.82 | −0.41 |
| 0.509 | 0.238 | 0.0720 | 0.101 | 0.0800 | 1.31 | 1.92 | −0.61 |
| 0.442 | 0.383 | 0.0290 | 0.0770 | 0.0690 | 1.21 | 2.10 | −0.89 |
| 0.523 | 0.262 | 0.0420 | 0.125 | 0.0480 | 1.23 | 2.06 | −0.83 |
| 0.446 | 0.330 | 0.0460 | 0.122 | 0.0560 | 1.28 | 2.0 | −0.72 |
| 0.346 | 0.0520 | 0.429 | 0.0960 | 0.0770 | 1.31 | 1.95 | −0.64 |
| 0.412 | 0.117 | 0.267 | 0.0960 | 0.108 | 1.43 | 1.79 | −0.36 |
| 0.426 | 0.466 | 0.00700 | 0.0560 | 0.0450 | 1.06 | 2.52 | −1.46 |
| 0.499 | 0.195 | 0.114 | 0.0950 | 0.0970 | 1.37 | 1.84 | −0.47 |
| 0.452 | 0.373 | 0.0270 | 0.0550 | 0.0930 | 1.20 | 2.14 | −0.94 |
| 0.327 | 0.0850 | 0.389 | 0.0800 | 0.119 | 1.39 | 1.84 | −0.45 |
| 0.414 | 0.129 | 0.234 | 0.158 | 0.0650 | 1.44 | 1.80 | −0.36 |
| 0.462 | 0.175 | 0.158 | 0.0830 | 0.122 | 1.42 | 1.79 | −0.37 |
| 0.323 | 0.0730 | 0.409 | 0.129 | 0.0660 | 1.37 | 1.88 | −0.51 |
| 0.432 | 0.443 | 0.0100 | 0.0780 | 0.0370 | 1.09 | 2.42 | −1.33 |
| 0.495 | 0.323 | 0.0310 | 0.0870 | 0.0630 | 1.20 | 2.10 | −0.90 |
| 0.423 | 0.158 | 0.204 | 0.0830 | 0.132 | 1.46 | 1.76 | −0.30 |
| 0.446 | 0.115 | 0.238 | 0.116 | 0.0850 | 1.41 | 1.80 | −0.39 |
| 0.458 | 0.166 | 0.168 | 0.120 | 0.0880 | 1.42 | 1.79 | −0.37 |
| 0.499 | 0.250 | 0.0680 | 0.109 | 0.0740 | 1.31 | 1.92 | −0.61 |
| 0.486 | 0.340 | 0.0250 | 0.0940 | 0.0550 | 1.19 | 2.15 | −0.96 |
| 0.455 | 0.166 | 0.176 | 0.0960 | 0.107 | 1.43 | 1.78 | −0.35 |
| 0.459 | 0.249 | 0.0970 | 0.0980 | 0.0970 | 1.39 | 1.83 | −0.44 |

$$(.1)$$

# Appendix B

With the notation introduced in Section 5.2, define the g × g matrix

$$(T_\beta)_{\sigma\tau} = n_\beta \text{tr} \, [\beta(\sigma^{-1}\tau)]/g,$$

in which the rows and columns are indexed by the elements of G. Let also

$$\mathcal{P}_\beta = n_\beta \sum_{\tau \in G} \chi_\beta(\tau^{-1})\phi(\tau)/g$$

indicate be the regular canonical projection associated with $\beta \in \widehat{G}$.

**Proposition B.1** $\mathcal{P}_\beta = T_\beta$.

**Proof:** To see this, use the fact that the single unitary entry in row $\sigma$ of $\phi(\gamma)$ is at column $\tau$ if and only if $\sigma = \gamma\tau$, or $\gamma = \sigma\tau^{-1}$, so that $(\mathcal{P}_\beta)_{\sigma\tau} = n_\beta \chi_\beta(\sigma\tau^{-1})/g = n_\beta \text{tr} \, \beta(\sigma\tau^{-1})/g = (T_\beta)_{\sigma\tau}$, as stated.

□

Moreover, because the Fourier transform $\widehat{x}(\beta) = \sum_{\tau \in G} x(\tau)\beta(\tau)$ of the scalar function x is also an element in $\mathbb{C}^{n_\beta \times n_\beta}$, the indexing $x_\beta(\tau) = n_\beta \mathrm{tr}\,[\beta(\tau^{-1})\widehat{x}(\beta)]/g$ is well-defined. The following proposition shows that it reduces as $\beta$.

**Proposition B.2** $x_\beta(\tau) = n_\beta \mathrm{tr}\,[\beta(\tau^{-1})\widehat{x}(\beta)]/g$ is an invariant of $\mathcal{P}_\beta$.

**Proof:** Using the identity $\mathcal{P}_\beta = T_\beta$ derived above, for an arbitrary vector $x \in \mathbb{R}^g$, it follows that

$$(\mathcal{P}_\beta x)_\sigma = \sum_\tau (T_\beta)_{\sigma\tau} x(\tau) = n_\beta \sum_\sigma \mathrm{tr}\,[\beta(\sigma^{-1})\beta(\tau)] x(\tau)/g$$

$$= n_\beta \mathrm{tr}\,[\beta(\sigma^{-1}) \sum_\tau x(\tau)\beta(\tau)]/g = n_\beta \mathrm{tr}\,[\beta(\sigma^{-1})\widehat{x}(\beta)]/g = x_\beta(\sigma).$$

That is, the entry $x(\tau)$ of the regular $\beta$-invariant $\mathcal{P}_\beta x$ is precisely $x_\beta(\tau) = n_\beta \mathrm{tr}\,[\beta(\tau^{-1})\widehat{x}(\beta)]/g$.

$\square$

**References**

[1]   A. I. Khinchin, *Information theory*, Dover, New York, NY, 1957.

[2]   S. Kullback, *Information theory and statistics*, Dover, New York, 1968.

[3]   Jean-Pierre Serre, *Linear representations of finite groups*, Springer-Verlag, New York, 1977.

[4]   M. Viana, *Symmetry studies- an introduction*, IMPA Institute for Pure and Applied Mathematics Press, Rio de Janeiro, Brazil, 2003.

[5]   _____, *Lecture notes on symmetry studies*, Technical Report 2005-027, EURANDOM, Eindhoven University of Technology, Eindhoven, The Netherlands, 2005. Electronic version http://www.eurandom.tue.nl.

[6]   M. L. Eaton, *Multivariate statistics- a vector space approach*, Wiley, New York, NY, 1983.

[7]   R. J. Muirhead, *Aspects of multivariate statistical theory*, Wiley, New York, 1982.

[8]   V. Lakshminarayanan and M. Viana, *Dihedral representations and statistical geometric optics I: Spherocylindrical lenses*, J. Optical Society of America A **22** (2005), no. 11, 2483–89.

[9]   M. Viana, *Invariance conditions for random curvature models*, Methodology and Computing in Applied Probability **5** (2003), 439–453.

[10]   M. Viana and V. Lakshminarayanan, *Dihedral representations and statistical geometric optics II: Elementary instruments*, J. Modern Optics- To appear (2006).

[11]   E. J. Hannan, *Group representations and applied probability*, Journal of Applied Probability **2** (1965), 1–68.

[12]   A. T. James, *The relationship algebra of an experimental design*, Annals of Mathematical Statistics **28** (1957), 993–1002.

[13]   U. Grenander, *Probabilities on algebraic structures*, Wiley, New York, NY, 1963.

[14]   R. A. Bailey, *Strata for randomized experiments*, Journal of the Royal Statistical Society B (1991), no. 53, 27–78.

[15] A. P. Dawid, *Symmetry models and hypothesis for structured data layouts*, Journal of the Royal Statistical Society B (1988), no. 50, 1–34.

[16] P. Diaconis, *Group representation in probability and statistics*, IMS, Hayward, California, 1988.

[17] M. L. Eaton, *Group invariance applications in statistics*, IMS-ASA, Hayward, California, 1989.

[18] R. H. Farrell (ed.), *Multivariate calculation*, Springer-Verlag, New York, NY, 1985.

[19] R. A. Wijsman, *Invariant measures on groups and their use in statistics*, Vol. 14, IMS, Hayward, California, 1990.

[20] L. Nachbin, *The Haar integral*, Van Nostrand, Princeton, N.J., 1965.

[21] S. Andersson, *Invariant normal models*, Annals of Statistics **3** (1975), no. 1, 132–154.

[22] M. D. Perlman, *Group symmetry covariance models*, Statistical Science **2** (1987), 421–425.

[23] M. Viana and D. Richards (M. Viana and D. Richards, eds.), Vol. 287, American Mathematical Society, Providence, RI, 2001.

[24] F.L. Ferris 3rd, V. Freidlin, A. Kassoff, S.B. Green, and R.C. Milton, *Relative letter and position difficulty on visual acuity charts from the early treatment diabetic retinopathy study.*, Am J Ophthalmol **15** (1993), 735–40.

[25] J. Aitchison, *The statistical analysis of compositional data*, Chapman and Hall, New York, NY, 1986.

[26] H. Doi, *Importance of purine and pyrimidine content of local nucleotide sequences (six bases long) for evolution of human immunodeficiency virus type 1*, Evolution **88** (1991), no. 3, 9282–9286.

[27] Y.M.M. Bishop, S.E. Fienberg, and P.W. Holland, *Discrete multivariate analysis: Theory and practice*, MIT Press, Cambridge, Massachusetts, 1975.