PROJECTION PURSUIT FOR DISCRETE DATA

BY

PERSI DIACONIS

TECHNICAL REPORT NO. 198
APRIL 1983

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION GRANT MCS80-24649

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
PROJECTION PURSUIT FOR DISCRETE DATA

Persi Diaconis
Department of Statistics
Stanford University

Abstract

Projection pursuit is an approach to analyzing high dimensional data by finding "interesting" or "structured" projections. This paper offers a definition of projections for discrete data. Theorems are proved to show that for most data sets, most projections are close to uniform. This suggests that projections are interesting if they partition data in a non-uniform way. The techniques suggested are applied to classify the books of Plato.

Key Words: Projection pursuit, Radon transform, Block designs, Exploratory data analysis.
INTRODUCTION

Multivariate data is often inspected by using linear projections. A number of interactive data-analysis programs allow projection of a high dimensional data-set into a low dimensional subspace selected by the user, who can search for interesting projections. Recently, Kruskal (1969, 1972), Friedman and Tukey (1979), and Huber (1981) have suggested computer "projection pursuit" algorithms for finding interesting projections.

This paper presents a notion of projection suitable for discrete-multivariate data. Definitions and examples are given in Section 1. The idea is to consider a partition of the underlying data space. The number of data points in each piece of the partition constitutes the projection. Varying the partition gives different projections. Section one presents conditions under which all projections determine the data set; developing a discrete version of the Radon transform. Section 2 gives a data analytic example in some detail. The data concerns the dating of the books of Plato. Projection pursuit leads to the discovery of a striking, easily interpretable structure that does not appear in other analysis of this data.

Section 3 proves that for most data sets, most partitions lead to approximately uniform projections. This leads directly to a useable criteria: a projection is interesting if it is far from uniform. The distance to uniformity can be measured by any distance between probabilities. The variation distance or entropy are natural candidates.

The final section gives results for the least uniform projection. Theorem 6 shows that if the class of projections is not too rich - for example, the affine hyperplane in $\mathbb{Z}^k$ - then for most data sets even the least uniform partition is close to uniform. If the class of projections contains many sets, then least uniform projections are "structured". The final theorem attacks the problem of a data analyst finding "structure" in "noise".
It may be useful to relate the results of this paper to the growing literature on Euclidean projection pursuit. Heuristically, a projection (as presented in a histogram or scatterplot) will be uninteresting if it is random or unstructured. One standard indication of randomness if large entropy. This gives a numerical criteria suggested by Huber (1981): a projection is interesting if it has small entropy relative to other projections. Huber observed that the numerical "measure of interest" used by Friedman and Tukey essentially minimizes \(-\int r^2\). This is a measure of entropy, quite similar to the more usual \(-\int f \log f\). If the scale is fixed, maximum entropy is attained at the Gaussian distribution. This suggests another heuristic: a projection is interesting if it is far from Gaussian. See Donoho (1981) for applications of this heuristic to time series. Diaconis and Freedman (1982) present a different rationale for looking at non-Gaussian projections. They showed that for many data sets, most projections have the same approximately Gaussian distribution. The results in sections 3 and 4 of the present paper give similar theorems for discrete data and the new notion of projection.

Projection pursuit was originally introduced as a way of selecting interesting projections. The idea gave rise to a battery of other techniques for doing high dimensional, non-linear, nonparametric curve fitting. These have mainly been developed by Friedman and Stuetzle (1981), (1982a,b). See Huber (1981) or Diaconis and Shahshahani for a review. Such techniques are useful in adjusting for observed structure. Most of the techniques seem to work in a straightforward way with discrete data and the present notion of projections. A more thorough study is certainly called for.

Acknowledgement. I thank Ethan Bolker, Victor Cullinan, and Shlomo Sternberg for making available their unpublished work on Radon transforms.
1. PROJECTIONS AND RADON TRANSFORMS

Let $X$ be a finite set. Let $Y$ be a class of subsets of $X$. Let $f : X \to \mathbb{R}$ be a function. The Radon transform of $f$ at $y \in Y$ is defined by

$$\mathcal{R}(y) = \int_{x \in y} f(x) .$$

The class $Y$ is called a projection base if

(1.2) $|y|$ is constant for $y \in Y$ ($|y|$ denotes the cardinality of $y$).

(1.3) There is a partition $p_1, \ldots, p_j$ of $Y$ such that each $p_i$ is a partition of $X$.

For a partition $p$, the numbers $\{\mathcal{R}(y)\}_{y \in p}$ will be called the projection of $f$ in direction $p$. The sets in $Y$ may be thought of as "lines" in a geometry. If lines in the same partition are called parallel, then (1.3) corresponds to the Euclidean axiom: for every point $x \in X$ and every line $y \in Y$ there is a unique line $y^*$ parallel to $y$ such that $x \in y^*$. In the statistics literature, designs with property 1.3 are called "resolvable". Assumption (1.2) guarantees that projections are based on averages over comparable sets. Consider the following examples.

**Example 1.** $X = \mathbb{Z}_2^k$ - the set of binary $k$-tuples.

Here is a concrete example of a data set with this structure; L. Brandwood classified each sentence of Plato's Republic according to its last five syllables. These can run from all short (u) through all long (-). Identifying u with 1 and - with 0, each sentence is associated with a binary 5-tuple. As $x$ ranges over $\mathbb{Z}_2^5$, let $f(x)$ denote the proportion of sentences with ending $x$. The values of $f(x)$ are given in the first column of Table 1.
### Table 1

Percentage Distribution of Sentence Endings

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>u u u u u u</td>
<td>1.1</td>
<td>2.4</td>
<td>2.5</td>
<td>1.7</td>
<td>2.8</td>
<td>2.4</td>
</tr>
<tr>
<td>- u u u u</td>
<td>1.6</td>
<td>3.8</td>
<td>2.8</td>
<td>2.5</td>
<td>3.6</td>
<td>3.9</td>
</tr>
<tr>
<td>u - u u u</td>
<td>1.7</td>
<td>1.9</td>
<td>2.1</td>
<td>3.1</td>
<td>3.4</td>
<td>6.0</td>
</tr>
<tr>
<td>u u - u u</td>
<td>1.9</td>
<td>2.6</td>
<td>2.6</td>
<td>2.6</td>
<td>2.6</td>
<td>1.8</td>
</tr>
<tr>
<td>u u u - u</td>
<td>2.1</td>
<td>3.0</td>
<td>4.0</td>
<td>3.3</td>
<td>2.4</td>
<td>3.4</td>
</tr>
<tr>
<td>u u u u -</td>
<td>2.0</td>
<td>3.8</td>
<td>4.8</td>
<td>2.9</td>
<td>2.5</td>
<td>3.5</td>
</tr>
<tr>
<td>- - u u u</td>
<td>2.1</td>
<td>2.7</td>
<td>4.3</td>
<td>3.3</td>
<td>3.3</td>
<td>3.4</td>
</tr>
<tr>
<td>- u - u u</td>
<td>2.2</td>
<td>1.8</td>
<td>1.5</td>
<td>2.3</td>
<td>4.0</td>
<td>3.4</td>
</tr>
<tr>
<td>- u u - u</td>
<td>2.8</td>
<td>0.6</td>
<td>0.7</td>
<td>0.4</td>
<td>2.1</td>
<td>1.7</td>
</tr>
<tr>
<td>- u u u -</td>
<td>4.6</td>
<td>8.8</td>
<td>6.5</td>
<td>4.0</td>
<td>2.3</td>
<td>3.3</td>
</tr>
<tr>
<td>u - - u u</td>
<td>3.3</td>
<td>3.4</td>
<td>6.7</td>
<td>5.3</td>
<td>3.3</td>
<td>3.4</td>
</tr>
<tr>
<td>u - u - u</td>
<td>2.6</td>
<td>1.0</td>
<td>0.6</td>
<td>0.9</td>
<td>1.6</td>
<td>2.2</td>
</tr>
<tr>
<td>u - u u -</td>
<td>4.6</td>
<td>1.1</td>
<td>0.7</td>
<td>1.0</td>
<td>3.0</td>
<td>2.7</td>
</tr>
<tr>
<td>u u - - u</td>
<td>2.6</td>
<td>1.5</td>
<td>3.1</td>
<td>3.1</td>
<td>3.0</td>
<td>3.0</td>
</tr>
<tr>
<td>u u u - -</td>
<td>4.4</td>
<td>3.0</td>
<td>1.9</td>
<td>3.0</td>
<td>3.0</td>
<td>2.2</td>
</tr>
<tr>
<td>u u u u -</td>
<td>2.5</td>
<td>5.7</td>
<td>5.4</td>
<td>4.4</td>
<td>5.1</td>
<td>3.9</td>
</tr>
</tbody>
</table>

| - - - u u       | 2.9  | 4.2  | 5.5   | 6.9  | 5.2   | 3.0  |
| - - u - u       | 3.0  | 1.4  | 0.7   | 2.7  | 2.6   | 3.3  |
| - - u u -       | 3.4  | 1.0  | 0.4   | 0.7  | 2.3   | 3.3  |
| - u - - u       | 2.0  | 2.3  | 1.2   | 3.4  | 3.7   | 3.3  |
| - u u - u       | 6.4  | 2.4  | 2.8   | 1.8  | 2.1   | 3.0  |
| - u u u -       | 4.2  | 0.6  | 0.7   | 0.8  | 3.0   | 2.8  |
| u u u u u       | 2.8  | 2.9  | 2.6   | 4.6  | 3.4   | 3.0  |
| u u u u -       | 4.2  | 1.2  | 1.3   | 1.0  | 1.3   | 3.3  |
| u u u u u       | 4.8  | 8.2  | 5.3   | 4.5  | 4.6   | 3.0  |
| u - - - u u      | 2.4  | 1.9  | 5.3   | 2.5  | 2.5   | 2.2  |

| u u u u u       | 3.5  | 4.1  | 3.3   | 3.8  | 2.9   | 2.4  |
| u u u u -       | 4.0  | 3.7  | 3.3   | 4.9  | 3.5   | 3.0  |
| u u u - u       | 4.1  | 2.1  | 2.3   | 2.1  | 4.1   | 6.4  |
| u u - u u       | 4.1  | 8.8  | 9.0   | 6.8  | 4.7   | 3.8  |
| u u - - u u      | 2.0  | 3.0  | 2.9   | 2.9  | 2.6   | 2.2  |

| - - - - u u      | 4.2  | 5.2  | 4.0   | 4.9  | 3.4   | 1.8  |

Number of sentences: 3,778, 3,783, 958, 770, 919, 762
A second example of data with this structure is the result of grading correct/incorrect in a test with \( k \) questions. There are several useful choices of \( Y \):

1a. **Marginal projections in** \( \mathbb{Z}_2^k \). For \( i = 1, 2, \ldots, k \), let \( y_i^0 = \{ x \in \mathbb{Z}_2^k : x_i = 0 \} \), let \( y_i^1 = \{ x \in \mathbb{Z}_2^k : x_i = 1 \} \). The sets \( Y = \{ y_i^j \}, 1 \leq i \leq 5, j \in \{0,1\} \), form a projection base. In the Plato example the projections have a simple interpretation as the proportion of sentences with a specific ending in the \( i \)th place. Displaying projections offers no problem here; a single number suffices.

A second natural choice of \( Y \) gives second order margins. This is based on sets \( y_{ij}^{ab} = \{ x \in \mathbb{Z}_2^k : x_i = a, x_j = b \} \), \( 1 \leq i < j \leq k \), \( a, b \in \{0,1\} \). In this case, a projection consists of 4 numbers. In the Plato example the projection along coordinates \( i,j \) gives the proportion of sentences with each of the 4 possible patterns \( uu, u-, -u, -- \) in positions \( i,j \). Table 3 in Section 2 is an example of one method of displaying such projections. Section 2 contains an analysis of the data in Table 1 based on these projections. The analysis gives a clear interpretation to a classical way of dating the books of Plato. The analysis is independent of the other examples in this section and can be read at this time.

1b. **Affine hyperplanes in** \( \mathbb{Z}_2^k \). This is one natural way of "filling out" the marginal projections in example 1a. For \( z \in \mathbb{Z}_2^k \) and \( a \in \{0,1\} \), let \( y_z^a = \{ x \in \mathbb{Z}_2^k : x \cdot z = a(\text{mod } 2) \} \). The sets \( Y = \{ y_z^a \}_{z \in \mathbb{Z}_2^k, a \in \{0,1\}} \) form a projection base. Observe that when \( z \) has a 1 in position \( i \) and zeros elsewhere, \( y_z^a \) equals the \( y_i^a \) of (1a). The sets in \( Y \) are the affine hyperplanes in \( \mathbb{Z}_2^k \).

Similarly the affine planes of any dimension form a projection base.

Here are some examples to show how the structure of \( f \) is reflected in \( \overline{f} \).

If \( f(x) = \delta_{x,x_0} \), \( \overline{f}(y) = 1 \) if \( x_0 \in y \) and zero otherwise. If \( f(x) = 1/2^k \), \( \overline{f}(y) = |y|/2^k \) for all \( y \). As a final example consider a fixed, non-zero vector \( y^* \in \mathbb{Z}_2^k \). Let \( S \) be the hyperplane determined by \( y^*: S = \{ x \in \mathbb{Z}_2^k : x \cdot y = 0(\text{mod } 2) \} \). Let
\[ f(x) = \begin{cases} 1/2^{k-1} & \text{for } x \in S \\ 0 & \text{otherwise.} \end{cases} \]

An easy computation shows

\[ \mathcal{F}(y_{z}^{c}) = \begin{cases} 1 & \text{if } z = y^* \\ \frac{1}{2} & \text{otherwise.} \end{cases} \]
\[ \mathcal{F}(y_{z}^{+}) = \begin{cases} 0 & \text{if } z = y^* \\ \frac{1}{2} & \text{otherwise.} \end{cases} \]

The hyperplane transform is essentially the same as the ordinary Fourier transform on the group \( \mathbb{Z}^{k}_2 \). This is defined by

\[ \hat{f}(z) = \sum_{x} (-1)^{x \cdot z} f(x). \]

If \( f \) is a probability on \( \mathbb{Z}^{k}_2 \), \( \hat{f}(z) = 2\mathcal{F}(y_{z}^{c}) - 1 \). The transform \( \hat{f} \) has been widely used for data analysis of this type of data. See Solomon (1961) or Diaconis (1983, Chapter 11).

The natural extension of this example to contingency tables will be treated in a separate paper.

**Example 2.** \( X = S_n \) - the set of permutations of \( n \) letters.

Permutation data arises in taste testing, ranking, and elections: for example, in presidential elections of the American Psychological Association, members are asked to rank order 5 candidates. Here, for a permutation \( \pi \), \( f(\pi) \) is taken as the proportion of voters choosing the order \( \pi \).

2a. Marginal projections. Let \( y_{ij} = \{\pi \in S_n : \pi(i) = j\} \) \( 1 \leq i, j \leq n \). These sets form a projection base. For fixed \( i \), the sets \( y_{i1}, y_{i2}, \ldots, y_{in} \) form a partition \( p(i) \). The projection in direction \( p(i) \) has a natural interpretation in the example: how did people rank candidate \( i \)? The projection can be displayed by making a histogram.
A second useful choice of $Y$ is based on considering two positions:

$$y_{ij}^{kl} = \{\pi \in S_n : \pi(i) = k, \pi(j) = l\} \quad i \neq j, k \neq l.$$  

This leads to projections giving the joint rankings of a fixed pair of candidates in the example. Such projections can be displayed by making a 2-dimensional picture and gray scaling the $(i,j)$ square to correspond to the proportion of voters ranking the pair of candidates in order $(i,j)$. Similarly third and higher order projections can be defined.

2b. Partitions based on subgroups. When $X$ is a group such as $S_n$, the following constructions for $Y$ are available. Let $N$ be a subgroup of $X$.

The orbits of $N$ acting on $X$ are the sets $\{Nx\}_{x \in X}$, the distinct orbits partition $X$. Varying by conjugation: $\{xNx^{-1}\}_{x \in X}$ gives a projection base for $X$.

When $N$ is taken as $S_{n-1} = \{\pi \in S_n : \pi(1) = 1\}$ the projections are the marginal projections defined in (2a) above. Taking $N$ as $S_{n-2} = \{\pi \in S_n : \pi(1) = 1, \pi(2) = 2\}$ gives the second order margins. An important class of subgroups are the so-called Young subgroups: let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ be a partition of $n$ so $\sum \lambda_i = n$. Let $S_{\lambda_1} \times S_{\lambda_2} \times \ldots \times S_{\lambda_n}$ be the permutations that permute the first $\lambda_1$ elements among themselves and the next $\lambda_2$ elements among themselves, etc. These include the previous examples and provide enough transforms for an inversion theory, as will be shown below. Display of such projections is not a well studied problem. In the case of a projection corresponding to a young subgroup, one suggestion is a 1-dimensional histogram using one of the orderings suggested in Chapter 3 of James (1978).

If $X = G/H$ where $G$ is a group and $H$ is a normal subgroup and $G \triangleright N \triangleright H$, with $N$ a subgroup, then the orbits of $N$ in $X$ are a partition and the orbits of $\{gNg^{-1}\}_{g \in G}$ form a partition base. One approach to the display of such projections is a 1-dimensional histogram using the ordering given by one of the metrics suggested in Chapter 7 of Diaconis (1983).
Example 3. $X = \mathbb{R}^p$ Euclidean data.

Consider data vectors $x_1, x_2, \ldots, x_n \in \mathbb{R}^p$. For $\gamma$ in the p-dimensional unit sphere, the projection in direction $\gamma$ is just $\gamma^*x_1, \ldots, \gamma^*x_n$. This is the classical Radon transform, with $Y$ consisting of the affine hyperplanes $y_t = \{x \in \mathbb{R}^p : x^*\gamma = t\}$. For fixed $\gamma$ these partition the space $\mathbb{R}^p$ as $t$ varies, and the partitions vary as $\gamma$ varies. In statistical applications a histogram is made of $\{\gamma^*x_i\}$ and one varies $\gamma$, trying to understand the structure of the p-dimensional data from the varying histograms.

Example 4. $X$ is a finite set with $n$ elements, $Y$ is the class of $k$-element subsets.

In this example, it is a non-trivial theorem of Baranyai that $Y$ forms a projection base. Details and discussion may be found in Cameron (1976). This example occurs naturally when considering extensions of a given class of partitions. For example, consider the marginal projections $y^a_1$ in $\mathbb{Z}_2^k$ defined in example 1a. These sets all have cardinality $|y^a_1| = 2^{k-1}$. It is natural to consider the extension to projections based on the class of all subsets of cardinality $2^{k-1}$.

Uniqueness of Radon Transforms. When is $f \leftrightarrow \tilde{f}$ one to one? A convenient criteria involves the notion of a block design. Let $|X| = n$. The class of sets $Y$ is a block design with parameters $(n,c;k,l)$ provided

(1.4) $|y| = c$ for all $y \in Y$

(1.5) each $x \in X$ is contained in $k$ subsets $y$

(1.6) each pair $x \neq x'$ is contained in $\ell$ subsets $y$.

Affine planes in $\mathbb{Z}_2^k$ and $k$ sets of an $n$ set (Example 4) are block designs. A great many other examples are discussed in the literature of combinatorial designs. In the statistics literature they are sometimes called balanced incomplete block designs. In the combinatorial literature they are often called
2-designs, or $2-(n,c,\lambda)$ designs. It is easy to see that the parameters $n, c, k, \lambda$ satisfy

\begin{align}
|Y|c &= n \cdot k \\
(n-1)\lambda &= k(c-1)
\end{align}

Demboski (1968) and Lander (1982) are useful references for block designs.

The following result is well known in the theory of design. I first learned of it from Bolker (1979) and Guilliman-Sternberg (1979).

**Theorem 1.** If $X$ is a finite set and $Y$ is a block design with $|Y| > 1$, then the Radon transform $f \mapsto \overline{f}$ is one to one, with an explicit inversion formula given by (1.11).

**Proof.** For any $x$,

\begin{align}
\sum_{x \in Y} \overline{f}(y) &= k f(x) + \lambda \sum_{x' \neq x} f(x') \\
&= (k-\lambda) f(x) + \lambda \sum_{y} f(x).
\end{align}

If $\Sigma f(x) = 1$, this determines $f$ as

\begin{align}
f(x) = \frac{1}{k-\lambda} \sum_{x \in Y} \overline{f}(y) - \frac{\lambda}{k-\lambda}.
\end{align}

Observe that $k > \lambda$ follows from the assumption that $|Y| > 1$. When $\Sigma f(x)$ is not known, it can be recovered by summing both sides of (1.9) in $x$. This gives

\begin{align}
\sum_{x} f(x) &= \frac{c}{k-\lambda} \sum_{y} \overline{f}(y) \\
\text{and so the inversion formula}
\end{align}

\begin{align}
f(x) = \frac{1}{k-\lambda} \sum_{x \in Y} \overline{f}(y) + \frac{\lambda c}{(k-\lambda)^2 + n \lambda(k-\lambda)} \sum_{y} \overline{f}(y).
\end{align}
Remarks. (a) It is not necessary that \( Y \) be a block design for \( f \to \tilde{f} \) to be one-to-one. For example, Kung (1979) shows that the Radon transform is one to one when \( Y \) consists of the sets of rank 1 in a matroid. Diaconis and Graham (1983) give examples where the transform is one to one when \( Y \) consists of the nearest neighbors in a metric space. For example, when \( X = \mathbb{Z}_2^{2k} \) and \( Y \) consists of the balls of Hamming distance less than or equal to 1 , the transform is one to one.

(b) The transform can still be useful and interesting if it is not one to one. For example, the marginal projections in examples 1a and 2a do not capture all aspects of the data but are often the first things to be looked at. In \( \mathbb{Z}_2^k \), if high enough marginal distributions are considered the function, \( f \) can be completely recovered. In the symmetric group, the projections corresponding to all Young subgroups determine \( f \) because they determine it's Fourier transform. See Diaconis (1983) for details.

2. DATA ANALYSIS OF SYLLABLE PATTERNS IN THE WORKS OF PLATO

This section presents a new analysis of data arising from syllable patterns in the works of Plato. The data is given in Table 1. It records, for each of 6 books, the pattern of long (\( - \)) and short (\( u \)) syllables among the last 5 syllables in each sentence. It is known that Plato wrote Republic early and Laws late. The other books were written between these but it is not known in what order. The goal of the analysis is to try to order the books. Our approach will be to study the books one at a time, trying to find patterns.

Projection pursuit suggests looking at various partitions of the data, searching for structured partitions which are far from uniform. Using first and second order margins as partitions, a reasonably striking difference between Republic and Laws is observed. This suggests a simple, interpretable way of ordering the other books as
Republic < Timaeus < Sophist < Politicus < Philebus < Laws.

This agrees with the standard ordering as discussed in Brandwood (1976, pg. XViii). Other analyses of this data set are in Cox and Brandwood (1959), Atkinson (1970), Wishart and Leach (1970) and Boneva (1971). The last reference contains a history and explanation for the choice of data. The first 3 analyses all use statistical models. Boneva's analysis uses a form of scaling. None of the previous analyses seem to have picked up the simple, striking pattern in the data that projection pursuit leads to.

The analysis is presented below, in a somewhat discursive style, in the order it was actually performed: first looking at the Republic, then Laws, and finally at the other books.

2a. The Republic. Table 2 shows the first order margins; e.g., the proportion of sentences with \( u \) in position \( 1, 1 \leq i \leq 5 \).

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion of ( u )</td>
<td>.465</td>
<td>.472</td>
<td>.466</td>
<td>.511</td>
<td>.363</td>
</tr>
</tbody>
</table>

Roughly, positions 1-4 are evenly divided between long and short. The last position is clearly different.

Table 3 lists the 10 possible pairs of second order margins.
Table 3 Republic

<table>
<thead>
<tr>
<th>uu</th>
<th>u-</th>
<th>-u</th>
<th>--</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,2)</td>
<td>.194</td>
<td>.271</td>
<td>.278</td>
</tr>
<tr>
<td>P (1,3)</td>
<td>.208</td>
<td>.257</td>
<td>.258</td>
</tr>
<tr>
<td>O (1,4)</td>
<td>.238</td>
<td>.227</td>
<td>.273</td>
</tr>
<tr>
<td>S (1,5)</td>
<td>.177</td>
<td>.288</td>
<td>.186</td>
</tr>
<tr>
<td>I (2,3)</td>
<td>.209</td>
<td>.263</td>
<td>.257</td>
</tr>
<tr>
<td>T (2,4)</td>
<td>.242</td>
<td>.230</td>
<td>.269</td>
</tr>
<tr>
<td>I (2,5)</td>
<td>.163</td>
<td>.309</td>
<td>.200</td>
</tr>
<tr>
<td>O (3,4)</td>
<td>.211</td>
<td>.255</td>
<td>.300</td>
</tr>
<tr>
<td>N (3,5)</td>
<td>.170</td>
<td>.296</td>
<td>.193</td>
</tr>
<tr>
<td>(4,5)</td>
<td>.168</td>
<td>.343</td>
<td>.195</td>
</tr>
</tbody>
</table>

A glance at the table shows that the first order effects are all too visible.

For example, the numbers in the first column (uu) are all "small" while the
numbers in the last column are "large". One simple way of adjusting for the
first order structure is to divide each number in Table 3 by the product of the
marginal totals. For example, in the first row, .194 would be divided by
.463 \times .472 (from Table 2) while .271 would be divided by .465 \times (1-.472).
The results are shown in Table 4:

Table 4 Republic

<table>
<thead>
<tr>
<th>uu</th>
<th>u-</th>
<th>-u</th>
<th>--</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,2)</td>
<td>.88</td>
<td>1.10</td>
<td>1.10</td>
</tr>
<tr>
<td>(1,3)</td>
<td>.96</td>
<td>1.03</td>
<td>1.03</td>
</tr>
<tr>
<td>(1,4)</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>(1,5)</td>
<td>1.05</td>
<td>.97</td>
<td>.96</td>
</tr>
<tr>
<td>(2,3)</td>
<td>.95</td>
<td>1.04</td>
<td>1.04</td>
</tr>
<tr>
<td>(2,4)</td>
<td>1.00</td>
<td>.99</td>
<td>1.00</td>
</tr>
<tr>
<td>(2,5)</td>
<td>.95</td>
<td>1.03</td>
<td>1.04</td>
</tr>
<tr>
<td>(3,4)</td>
<td>.89</td>
<td>1.12</td>
<td>1.10</td>
</tr>
<tr>
<td>(3,5)</td>
<td>1.00</td>
<td>1.00</td>
<td>.99</td>
</tr>
<tr>
<td>(4,5)</td>
<td>.91</td>
<td>1.05</td>
<td>1.10</td>
</tr>
</tbody>
</table>
Most of the ratios are close to 1, so a product model is a reasonable first description. The projection pursuit approach suggests that a partition of the data (here a row) is "interesting" if the partition is far from uniform. By eye, looking at Table 4, positions (1,2), (2,3), (3,4), (4,5) are far from being all 1. Observe that these positions are adjacent, as (i,i+1).

Next observe that each of the 4 designated rows has a common pattern: the first and last entries are small, the middle two entries are large. Going back to the definitions, this pattern arises from a negative association of adjacent syllables: in the Republic adjacent syllables tend to alternate. The pattern in row (1,3) argues that this cannot be a complete description; after all, if the symbols alternate, the position 2 apart should be positively associated, but (1,3) displays negative association. Looking at the other rows of the table, we observe that the size goes big, small, small, big or its opposite, small, big, big, small. This is an artifact. Consider the first row of Table 4. It was formed from 4 proportions that sum to 1, w, x, y, z say. The 4 adjusted entries are

\[
\begin{align*}
\frac{w}{(w+x)(w+y)} & \quad \frac{x}{(x+w)(x+z)} & \quad \frac{y}{(y+z)(y+w)} & \quad \frac{z}{(z+y)(z+x)}
\end{align*}
\]

It is easy to show that the first entry is less than 1 if and only if the second is larger than 1, if and only if the third is larger than one if and only if the fourth is less than 1. This means that the first column in Table 4, together with the first order margins, determines the remaining entries. This artifact in no way reflects on the association pattern noted earlier - the most structured rows correspond to adjacent syllables, and adjacent syllables are negatively associated.
2b. **Laws** (and a comparison with the Republic). The first order margins for **Laws** are only slightly different from those in Republic

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion u</td>
<td>.477</td>
<td>.489</td>
<td>.411</td>
<td>.599</td>
<td>.375</td>
</tr>
</tbody>
</table>

The pattern is the same: overall, fewer than half u's; the last position sharply smaller. The similarity between the first order margins in Republic and **Laws** suggests that second or higher order margins must be used to order the remaining books. The analog of the first column of Table 4 is given below:

<table>
<thead>
<tr>
<th>(1,2)</th>
<th>(1,3)</th>
<th>(1,4)</th>
<th>(1,5)</th>
<th>(2,3)</th>
<th>(2,4)</th>
<th>(2,5)</th>
<th>(3,4)</th>
<th>(3,5)</th>
<th>(4,5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>uu</td>
<td>1.07</td>
<td>1.03</td>
<td>.92</td>
<td>.99</td>
<td>1.43</td>
<td>.98</td>
<td>.98</td>
<td>1.04</td>
<td>1.09</td>
</tr>
</tbody>
</table>

A typical entry being the proportion of sentences with uu in the given position divided by the product of the marginal proportions.

Again, pairwise adjacent positions are associated, all in the same way. Here, the association is positive, whereas for the Republic the association is negative. This is the striking pattern referred to above. It suggests a method of ranking the other books: compare the sign pattern or actual ratios of the adjusted second order margins of other books with the Republic and **Laws**. For definiteness, the sum of absolute deviations between the second order margins over all 10 positions will be used.

2c. Analysis for Philebus and Politicus. These books are somewhat similar to each other. The first and second order margins for **Philebus** are as follows:
Table 7 - Philebus

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proposition of u</td>
<td>.502</td>
<td>.464</td>
<td>.398</td>
<td>.594</td>
<td>.445</td>
</tr>
</tbody>
</table>

Table 8 - Philebus

<table>
<thead>
<tr>
<th>Position</th>
<th>(1,2)</th>
<th>(1,3)</th>
<th>(1,4)</th>
<th>(1,5)</th>
<th>(2,3)</th>
<th>(2,4)</th>
<th>(2,5)</th>
<th>(3,4)</th>
<th>(3,5)</th>
<th>(4,5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusted uu</td>
<td>1.15</td>
<td>1.07</td>
<td>.89</td>
<td>1.11</td>
<td>1.48</td>
<td>.92</td>
<td>.89</td>
<td>1.02</td>
<td>1.00</td>
<td>1.06</td>
</tr>
</tbody>
</table>

Note the difference in first order margins: between Philebus and Republic (or Laws) position 1 is high, as are positions 4 and 5. For second order margins, the adjacent patterns are all positively associated ((2,3) being truly extreme). Comparing Table 8 with Table 6, the association pattern matches Laws in direction, except in position (1,5).

The relevant averages for Politicus are:

Table 9 - Politicus

<table>
<thead>
<tr>
<th>Position</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proportion u</td>
<td>.477</td>
<td>.457</td>
<td>.348</td>
<td>.524</td>
<td>.469</td>
</tr>
</tbody>
</table>

Table 10 - Politicus

<table>
<thead>
<tr>
<th>Position</th>
<th>(1,2)</th>
<th>(1,3)</th>
<th>(1,4)</th>
<th>(1,5)</th>
<th>(2,3)</th>
<th>(2,4)</th>
<th>(2,5)</th>
<th>(3,4)</th>
<th>(3,5)</th>
<th>(4,5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusted uu</td>
<td>1.17</td>
<td>1.10</td>
<td>.96</td>
<td>1.01</td>
<td>1.26</td>
<td>.87</td>
<td>.90</td>
<td>1.05</td>
<td>1.10</td>
<td>1.13</td>
</tr>
</tbody>
</table>

The first order margins are, very roughly, like those in both the Republic and Laws, but again the 3rd position has a low proportion of short syllables. The second order margins have the same pattern as Laws. The same remarks made for the second order margins of Philebus apply.

Both Philebus and Politicus seem very similar to Laws. Which of these two is closest to Laws? One simple approach is to consider the sum of the absolute
values of the difference between the entries of Tables 8 and 6 along with the
difference between 10 and 6. The sum for Laws to Philebus is .64 while the
sum for Laws to Politicus is .83. Thus a tentative ranking is:

Politicus < Philebus < Laws.

2d. Analysis for Sophist and Timaeus. These books are quite similar to each
other and, as we shall see, quite different from Laws, Philebus, and Politicus.

<table>
<thead>
<tr>
<th>Table 11 - Sophist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
</tr>
<tr>
<td>Proportion u</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 12 - Sophist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
</tr>
<tr>
<td>Adjusted uu</td>
</tr>
</tbody>
</table>

The first order margins are quite different from the books examined previously.
They are, roughly, consistent with all syllables being equally likely to be long
or short. The first order pattern seems closest to Politicus. The second order
associations are closer to 1 than in Laws, Politicus, or Philebus. Adjacent
positions are positively associated, except for (3,4). The direction of associa-
tion matches Laws in only 6 of the 10 positions. The sum of absolute deviations
between the entries of Tables 6 and 12 is .87. Finally, turn to Timaeus.

<table>
<thead>
<tr>
<th>Table 13 - Timaeus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Position</td>
</tr>
<tr>
<td>Proportion u</td>
</tr>
</tbody>
</table>
Table 14 - Timaeus

<table>
<thead>
<tr>
<th>Position</th>
<th>(1,2)</th>
<th>(1,3)</th>
<th>(1,4)</th>
<th>(1,5)</th>
<th>(2,3)</th>
<th>(2,4)</th>
<th>(2,5)</th>
<th>(3,4)</th>
<th>(3,5)</th>
<th>(4,5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adjusted uu</td>
<td>1.01</td>
<td>1.02</td>
<td>.99</td>
<td>1.04</td>
<td>.94</td>
<td>.95</td>
<td>.99</td>
<td>.99</td>
<td>.98</td>
<td>1.08</td>
</tr>
</tbody>
</table>

A distinctive feature of the first order margins is the large proportion of short syllables in the third position. The adjusted second order margins are close to 1; so Timaeus seems closest to Sophist. Of the 4 adjacent positions, two show positive association and two show negative association. The direction of association matches Laws in 6 positions; the sum of absolute deviations between Tables 14 and 6 is .94. The distance between Timaeus and the Republic (Tables 14 and 4) is .60, so Timaeus seems closer to Republic than to Laws using this measure.

Because of the decrease in the number of matches, and the increase in the sum of absolute deviations, it seems reasonable to rank

Republic < Timaeus < Sophist.

This completes the discussion of this example.

3. MOST PROJECTIONS ARE UNIFORM

The theorems of this section imply that for most probabilities \( f(x) \), most projections \( \tilde{f}(y) \) are about the same - close to uniform. This necessitates projection pursuit - choosing projections that are far from uniformly distributed - to determine what is special about a particular \( f \). This gives an independent rational for a suggestion of Huber who has suggested that Euclidean projections are interesting if they are far from uniform in the sense of having minimum entropy (of course, the uniform distribution has maximum entropy).
Theorem 2. Let $X$ be a finite set with $n$ elements. Let $Y$ be a block design with block size $c$ (so $|y| = c$ for $y \in Y$). Let $f : X \to \mathbb{R}$ be any function and let $\mu(f) = \sum_{x \in X} f(x)$. Let $y$ be chosen uniformly in $Y$. Then

\begin{equation}
E[f(y)] = \frac{c}{n} \mu(f)
\end{equation}

(3.1)

\begin{equation}
\text{Var}[f(y)] = \frac{c}{n} \left( 1 - \frac{c-1}{n-1} \right) \mu(f - \frac{\mu(f)}{n})^2
\end{equation}

(3.2)

Proof for (3.1):

\[E[f(y)] = \frac{1}{|Y|} \sum_{y \in Y} f(y) = \frac{1}{|Y|} \sum_{y \in Y} \left( \sum_{x \in Y : y \subset x} \frac{1}{|Y|} \sum_{x \in Y : |y| x \subset y} f(x) \right) = \frac{k}{|Y|} \mu(f).
\]

From (1.7), $k/|Y| = c/n$.

For (3.2) assume, without loss of generality, that $\mu(f) = 0$. Then

\[\text{Var}(f(y)) = \frac{1}{|Y|} \sum_{y \in Y} (f(y))^2 = \frac{1}{|Y|} \left\{ \sum_{x \in Y} f(x)^2 + 2 \sum_{x \neq x'} f(x) f(x') \right\}
\]

\[= \frac{k}{|Y|} \mu(f^2) + \frac{2l}{|Y|} \sum_{x \neq x'} f(x) f(x') = \frac{k-l}{|Y|} \mu(f^2) + \frac{l}{|Y|} \mu(f^2)
\]

\[= \frac{k-l}{|Y|} \mu(f^2).
\]

From (1.7,8) $(k-l)/|Y| = c(n-c)/n(n-1)$.

Example a. When $Y$ is the $j$ sets of an $n$ set, $|Y| = \binom{n}{j}$, $c = j$, and the result reduces to the usual mean and variance for a sample without replacement.

Example b. Let $X = \mathbb{Z}_2^k$ and $Y$ = the $j$-dimensional affine planes. Then $n = 2^k$ and $c = 2^{k-j}$. If $\mu(f) = 1$, the results become

\[E[f(y)] = \frac{1}{2^j}, \quad \text{Var}[f(y)] = \frac{2^j}{2^k} \left( 1 - \frac{2^{j-1}}{2^{k-1}} \right) \mu(f - \frac{1}{2^k})^2.
\]
For future use, observe that the cardinality of \( Y \) in this case is

\[
\frac{2^j(2^{k-1}) \cdots (2^{k-2}) \cdots (2^{k-2j+1})}{(2^j-1) \cdots (2^j-2^{j-1})}.
\]

Returning to the situation in Theorem 1, Chebychev's inequality implies:

**Corollary 1.** With notation as in Theorem 1, the proportion of \( y \in Y \) such that

\[
|\bar{T}(y) - \frac{c}{n} \mu(f)| > \epsilon
\]

is smaller than

\[
\frac{1}{\epsilon^2} \frac{c}{n(1 - \frac{c-1}{n-1})} \mu(f - \frac{\mu(f)}{n})^2.
\]

**Remarks.** Thus, for functions \( f \) which are "not too wild" in the sense that \( \mu(f - \frac{\mu(f)}{n})^2 \) is small, most transforms \( \bar{T}(y) \) are uninformative in the sense of being close to their mean value. As an example, take \( X = \mathbb{Z}_2 \) and \( f \) the function defined by the first column of Table 1. Then \( \mu(f - \frac{1}{32})^2 = .0021. \) If \( Y \) is taken as the set of all affine hyperplanes, the corollary gives that 95% of the transforms have \( |\bar{T}(y) - 1/2| < .04. \)

The next theorem says that for most probabilities \( f, \mu(f - \frac{1}{n})^2 \) is small (about \( 1/n \)).

**Theorem 4.** Let \((U_1, U_2, \ldots, U_n)\) be chosen uniformly on the \( n \) simplex.

For large \( n \), the random variable

\[
\frac{n^{3/2}}{2} \left\{ \sum_{i=1}^{n} \frac{1}{n} (U_i - 1/n)^2 - 1/n \right\}
\]

has an approximate standard normal distribution.

**Proof.** The argument uses the representation of a uniform distribution by means of exponential variables. Let \( X_1, X_2, \ldots, X_n \) be independent standard exponential variables with density \( e^{-x} \) on \([0, \infty)\). Let
\[ S_1 = \frac{n}{\sqrt{n}} x_1, \quad S_2 = \frac{n}{\sqrt{n}} x_2^2. \]

For large \( n \), the random vector

\[
\begin{pmatrix}
    Z_1 \\
    Z_2
\end{pmatrix} = \frac{1}{\sqrt{n}} \begin{pmatrix} S_1 - n \ \ 4 \ \ 20 \end{pmatrix}
\]

has an approximate bivariate normal distribution with mean vector zero and covariance matrix \( \begin{pmatrix} 4 & 20 \\ 20 & 20 \end{pmatrix} \). To check the covariance matrix, check that

\[
\text{Var}\left(\frac{S_1 - n}{\sqrt{n}}\right) = \text{Var}(X_1) = 1, \quad \text{Var}\left(\frac{S_2 - n}{\sqrt{n}}\right) = \text{Var}(X_2) = 20; \quad \frac{1}{n} \text{E}\{(S_1 - n)(S_2 - 2n)\} = \text{E}\{X_1(X_2 - 2)\} = \text{E}(X_1^3) - \text{E}(X_1^2) - 2\text{E}(X_1) + 2 = 4.
\]

Now represent a uniform vector on the \( n \) simplex as \( U_1 = X_1/S_1 \). Then

\[
\Sigma(U_1 - 1/n)^2 = \frac{1}{S_1^2} \Sigma X_1^2 - \frac{1}{n} = \frac{1}{S_1^2} \Sigma(X_1^2 - 2) + \frac{2n}{S_1^2} - \frac{1}{n}.
\]

Now \( S_1 = n(1 + \frac{1}{\sqrt{n}} Z_1) \) with \( Z_1 = \frac{S_1 - n}{\sqrt{n}} \). Thus

\[
S_1^2 = n^2 \left(1 + \frac{2}{\sqrt{n}} Z_1 + \frac{4Z_1^2}{n^2}\right).
\]

Using the standard \( O_p \) notation

\[
\frac{1}{S_1^2} = \frac{1}{n^2} - \frac{2Z_1}{n^{3/2}} + O_p\left(\frac{1}{n^3}\right).
\]

Thus

\[
\frac{1}{S_1^2} \Sigma(X_1^2 - 2) = \frac{1}{n^{3/2}} \frac{1}{\sqrt{n}} \Sigma(X_1^2 - 2) + O_p\left(\frac{1}{n^{2}}\right) = \frac{1}{n^{3/2}} Z_2 + O_p\left(\frac{1}{n^2}\right),
\]

\[
\frac{2n}{S_1^2} = \frac{2}{n} - \frac{4Z_1}{n^{3/2}} + O_p\left(\frac{1}{n^2}\right).
\]

The bivariate limiting normality of \( \begin{pmatrix} Z_1 \\
Z_2
\end{pmatrix} \) implies that \( Z_2 = 4Z \), has an approximate normal distribution, with mean 0 and variance

\[
\text{Var}(Z_2) + 16\text{Var}(Z_1) - 8\text{Cov}(Z_1, Z_2) = 4.
\]

\[\square\]
Theorems 2 and 3 imply that for most probabilities $f$, most transforms $\tilde{f}(y)$ are close to uniform. The final result of this section deals with the entire projection $\{\tilde{f}(y)\}_{y \in p}$ where $p$ is a partition of $X$ into blocks in $Y$.

Let $X$ be a finite set. Let $Y$ be a block design on $X$ with parameters $(n, c, k, \ell)$. Suppose that $Y$ is also a projection base for $X$ with $p_1, p_2, \ldots, p_j$ being a partition of $Y$, each $p_i$ being a partition of $X$. Of course $J = |Y|c/n$. The next theorem implies that for most functions, a randomly chosen partition is uniformly close to $c/n$.

**Theorem 5.** Let $Y$ be a block design on $X$ with parameters $(n, c, k, \ell)$. Suppose that $Y$ is a projection base. Let $f$ be a fixed probability on $X$. Let the partition $p$ be chosen at random. For $\varepsilon > 0$,

\[(3.3) \quad \sum_{y \in p} |\tilde{f}(y) - \frac{c}{n}| < \varepsilon.\]

With probability at least

\[1 - \frac{n(n-c)}{c^2 c(n+1)} \mu(f - \frac{1}{n})^2 \cdot\]

**Proof of Theorem 5.** The probability model for choosing a random partition is based on a fixed enumeration $p_1, p_2, \ldots, p_j$ of the partitions that make up $Y$. Each partition is assumed to be taken in a fixed order $y_i^n = (y_1^i, \ldots, y_j^i n/c)$. The random variable $S(p) = \sum_{y \in p} |f(y) - c/n|$ is invariant under permuting the $y \in p$ among themselves. Thus a random variable with the same distribution of $S(p)$, but exchangeable $\{f(y)\}_{y \in p}$ exists. For this realization, $E(\sum_{y \in p} |f(y) - c/n|) = \frac{n}{c} E|f(y^*) - \frac{c}{n}|$, with $y^*$ chosen uniformly in $Y$. Using Cauchy-Schwartz and Theorem 2, the expectation is bounded above by

\[\frac{n}{c} \left( \frac{c-1}{n-1} \right) \mu(f - \frac{1}{n})^2.\]

Theorem 4 follows from this bound and Markov's inequality applies to the original random variable.
Remarks. From Theorem 3, \( \mu(f - \frac{1}{n})^2 \geq 1/n \) for most functions \( f \). For such \( f \), the theorem implies that for large block size \( c \), most partitions are close to uniform in variation distance. When \( c \) is small, most partitions are far from uniform in variation distance. This may be contrasted with Theorems 2 and 3 which imply that the components \( \hat{P}(y) \) of most projections are close to \( c/n \). When \( c \) is small, there are many terms in the sum (3.3). As an example, consider the 2-sets of an \( n \) set where \( n = 2^j \). Let \( p \) be a random partition into 2 element sets. Let \( f \) be chosen at random from the \( n \) simplex. It is straightforward to show that with probability tending to 1 as \( n \) tends to infinity

\[
\sum_{y \in p} |\hat{P}(y) - \frac{2}{n}| \rightarrow 8c^{-2}.
\]

It is natural to ask for a central limit theorem in connection with Theorems 2 and 4. For \( j \) sets of an \( n \)-set, such a theorem is available from the usual results on sampling without replacement from a finite population. Most likely, there is a similar set of results for block designs with \( |Y| \) and \( c \) large.

4. LEAST UNIFORM PARTITIONS

The results of section 3 imply that, under suitable conditions, for most functions the projection along most partitions is close to uniform. This suggests that the special properties of particular functions are only seen in partitions that are far from uniform. In this section, properties of least uniform partitions are examined. Theorem 6 shows that for most functions, even the least uniform partitions will be close to uniform if the number of sets in \( Y \) is small in the sense that \( \log Y \) is small compared to \( n \) and the block size \( c \). This is true, in particular, for the affine hyperplanes in \( \mathbb{Z}_2^k \).
Theorem 6. Let $X$ be a set of $n$ elements. Let $Y$ be a class of subsets in $X$ of fixed cardinality $c$. Suppose that $p_1, p_2, \ldots, p_j$ is a partition of $Y$ into partitions of $X$. Let $f$ be chosen at random in the $n$ simplex. Let $p^*$ be the partition in $\{p_i\}$ that maximizes $\sum_{y \in p} |\overline{f}(y) - c/n|$. For any $\epsilon > 0$

$$\sum_{y \in p^*} |\overline{f}(y) - c/n| < \epsilon,$$

Except perhaps for a set of $f$'s of probability smaller than

$$\left(|X| + 1\right) \beta$$

with $\beta$ equal to 1 minus

$$(4.1) \quad \frac{1}{\beta(c,n)} \int_{\frac{c}{n}(1-\epsilon)}^{\frac{c}{n}(1+\epsilon)} x^{c-1} (1-x)^{n-c-1} dx$$

where $\beta(c,n)$ denotes the beta function.

Proof. Represent a randomly chosen $f$ as $X_1/S$, where $X_1$ are independent standard exponentials and $S = \prod_{i=1}^n X_i$. Let $y^*$ be the set in $Y$ with the largest value of $\overline{f}(y)$. The argument begins by bounding the probability that

$$|\overline{f}(y^*) - c/n| < \epsilon \frac{c}{n}.$$ 

To begin with,

$$P(\overline{f}(y^*) < \frac{c}{n} (1-\epsilon)) \leq P\left(\frac{\sum X_i}{S} < \frac{c}{n} (1-\epsilon)\right)$$

further,

$$P(\overline{f}(y^*) > \frac{c}{n} (1+\epsilon)) \leq \sum_{y \in Y} P(\overline{f}(y) > \frac{c}{n} (1+\epsilon)) = |Y| P\left(\frac{\sum X_i}{S} > \frac{c}{n} (1+\epsilon)\right).$$

Next, let $y_*$ denote the set in $Y$ with the smallest value of $\overline{f}(y)$; to bound the probability that $|f(y_*) - c/n| < \epsilon c/n$, observe that
\( f(y_*) = 1 - \bar{f}(y^{**}) \) with \( y^{**} \) the union of sets in a partition omitting one element that maximizes \( \bar{f} \). Thus,

\[
P(\bar{f}(y_*) < \frac{c}{n} (1 - \varepsilon)) = P(\bar{f}(y^{**}) > 1 - \frac{c}{n} (1 - \varepsilon)) \leq |Y| \left| P\left( \frac{X_1 + \ldots + X_{n-c}}{s} > 1 - \frac{c}{n} (1 - \varepsilon) \right) \right|
\]

Further,

\[
P(\bar{f}(y_*) > \frac{c}{n} (1 + \varepsilon)) = P(f(y^{**}) < 1 - \frac{c}{n} (1 + \varepsilon)) \leq P\left( \frac{X_1 + \ldots + X_{n-c}}{s} < 1 - \frac{c}{n} (1 + \varepsilon) \right)
\]

Summing the 4 bounds thus obtained we see that both

\[(4.2) \quad |f(y_*) - \frac{c}{n}| < \varepsilon \frac{c}{n} \quad \text{and} \quad |f(y^{**}) - \frac{c}{n}| < \varepsilon \frac{c}{n} \]

except for a set of \( f \)'s of probability smaller than \((|Y| + 1)\beta\) as defined by (4.1). Now (4.2) implies that \(|f(y) - c/n| < \varepsilon \frac{c}{n}\) for all \( y \in Y \). Summing this last inequality over the partition \( p^* \) completes the proof of the theorem.\[\square\]

Remarks. The beta integral that appears in the bound is straightforward to approximate numerically. A raft of techniques and approximations appear in the first chapter of Pearson (1968). For example, consider cases where \( c/n = 1/2 \).

Then, using the Peiser-Pratt approximation given in Pearson, and Mills ratio, the \( \beta \) in (4.1) is approximately

\[
\frac{2}{\sqrt{2\pi}} \cdot e^{-x^2/2} \quad \text{with} \quad x = \left\{ 2c \log[1/(1/(k_\varepsilon))(1+c)] \right\}^{1/2}
\]

For this to be small when multiplied by \(|Y| + 1\), it clearly suffices that \( \log|Y| \) be small compared to \( c \). This is the case for the affine subspaces of dimension \( j \) in \( \mathbb{Z}_2^k \) if \( j \) is bounded and \( k \) is large.

As a numerical example, consider the affine hyperplanes in \( \mathbb{Z}_2^{10} \). Then \(|Y| + 1 = 2049, c = 512, n = 1024 \). Taking \( \varepsilon = .1, (|Y| + 1)\beta = 2.595 \times 10^{-7} \).
The next theorem shows that when there are many sets in $Y$, the least uniform projection is typically far from uniform. The theorem deals with $n$ sets in a set of cardinality $2n$. The variation distance of a typical probability projected along the least uniform half split is shown to be about $0.3$. This may be compared with Theorems 2 and 3 which show that for a typical probability $f$ on $2n$ points $|\bar{f}(y) - 1/2|$ is close to zero for most sets $y$ of cardinality $n$.

**Theorem 7.** Let $f$ be chosen at random on the $2n$ simplex. Let $S^-$ be the sum of the $n$ smallest $f(i)$. Then, for large $n$, the random variable

$$\sqrt{2n}(S^- - \left(\frac{1}{2} - \frac{\log 2}{2}\right))$$

has an approximate normal distribution with mean $0$ and variance $\frac{3}{2} - 2 \log 2$.

**Proof.** Represent a randomly chosen $f$ as $X_i/S$ where $X_i$ are independent standard exponential variables and $S = \sum_{i=1}^{2n} X_i$. Denote the order statistics by round brackets:

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(2n)}.$$

Let $L_1 = X_{(1)}$, $L_2 = X_{(2)} - X_{(1)}$, $\ldots$, $L_{2n} = X_{(2n)} - X_{(2n-1)}$. Then, the $L_i$ are independent, and $L_{i+1}$ has the distribution of a standard exponential times $1/(2n-i)$ - see Feller (1971). With this notation,

$$S = \sum_{i=1}^{2n} X_i = \sum_{i=0}^{2n-1} (2n-i)L_{i+1},$$

(4.3)

$$S^- = \frac{1}{S} \sum_{i=1}^{n} X_{(i)} = \frac{1}{S} \sum_{i=0}^{n-1} (n-i)L_{i+1}.$$

The proof is completed by approximating the sums in this representation of $S$ and $S^-$. Let $\mu_i = \frac{n-i}{2n-1}$, so $(n-i)L_{i+1}$ has the same distribution as $\mu_i$ times a standard exponential. Let
\[
\mu = \sum_{i=0}^{n-1} \mu_i = \sum_{i=0}^{n-1} \left(1 - \frac{n}{2n-i}\right) = n - n \log 2 + o(1).
\]

Let
\[
\sigma^2 = \sum_{i=0}^{n-1} \sigma_i^2 = \sum_{i=0}^{n-1} \left(1 - \frac{2n}{2n-i} + \frac{n^2}{(2n-i)^2}\right)
\]

\[= 2(n - [2n \log 2 + o(1)] + \left[\frac{3}{2} + o(1)\right])
\]

\[= 2n\left[\frac{3}{2} - 2 \log 2\right] + o(1).
\]

Let \(Z_1 = (S-2n)/\sqrt{2n}\) and \(Z_2 = (\Sigma_{i=1}^{n} X_{i} - \mu)/\sqrt{2n}\). The vector \((Z_1, Z_2)\) has a limiting bivariate normal distribution, with mean \((0,0)\) and covariance matrix \((\sigma_1^2 \rho \sigma_2^2)\) with \(\sigma_1^2 = 2, \sigma_2^2 = \frac{3}{2} - 2 \log 2\), and \(\rho = \frac{1}{2}(1 - \log 2)\). To check the value of \(\rho\), observe that the covariance of \(Z_1\) and \(Z_2\) is \(1/2n\) times

\[\sum_{i=0}^{n-1} E[(2n-i)L_{i+1} - 1][(n-i)L_{i+1} - \frac{n-i}{2n-i}],\]

\[= \sum_{i=0}^{n-1} \frac{n-i}{2n-i} = n - n \log 2 + o(1).
\]

Using the standard \(p\) calculus,

\[\frac{1}{S} = \frac{1}{2n} \frac{1}{(1+Z_1/\sqrt{2n})} = \frac{1}{2n} \left(1 - \frac{Z_1}{\sqrt{2n}}\right) + o_p\left(\frac{1}{n}\right).
\]

In particular,

\[\frac{1}{S} = \frac{1}{2n} + o_p\left(\frac{1}{n^{3/2}}\right).
\]

The representation (4.3) for \(S^2\) can be rewritten as

\[S^2 = \sqrt{2n} \frac{Z_2}{S} + \frac{\mu}{S} = \frac{Z_2}{\sqrt{2n}} + \frac{1 - \log 2}{2} \left(1 - \frac{Z_1}{\sqrt{2n}}\right) + o_p\left(\frac{1}{n}\right).
\]

It follows that \(\sqrt{2n}(S^2 - \frac{1 - \log 2}{2})\) has the same limiting distribution as \(Z_2 - (\frac{1 - \log 2}{2})Z_1\). This is normal, with mean 0 and variance

23
\[ \left( \frac{3}{2} - 2 \log 2 \right) + 2(\frac{1}{2} - \frac{1}{2} \log 2) - 2(\frac{1}{2} - \frac{1}{2} \log 2) = \frac{3}{2} - 2 \log 2. \]

**Corollary 2.** Let \( f \) be chosen at random on the \( 2n \) simplex. Let \( (y, y^c) \) be a partition of \( X \) into an \( n \) set and its complement which maximizes the value of

\[ |\bar{f}(y) - \frac{1}{2}| + |\bar{f}(y^c) - \frac{1}{2}|. \]

Then, as \( n \) tends to infinity, the maximum discrepancy tends to \( \log 2 \approx .301 \) with probability tending to 1.

**Proof.** For almost all \( f \), the maximum is taken on uniquely at the partition \( S^-, S^-^c \) as defined in Theorem 7. The maximum discrepancy equals

\[ 2|S^- - \frac{1}{2}|. \]

and the result follows from Theorem 7.

**Remark.** The proof of Theorem 7 and its corollary can easily be extended to cover the \( j \) sets of an \( n \) set. The argument shows that for most probabilities \( f \), the variation distance between the least uniform projection and the uniform distribution is bounded away from zero if \( j \) is an appreciable fraction of \( n \).

For the final theorem, a different method of choosing a random probability is introduced. Let \( X \) be a set of cardinality \( 2n \). Fix an integer \( b \). Drop \( b \) balls into \( 2n \) boxes, and let \( f(x) \) be the proportion of balls in the box labeled \( x \). Let \( Y \) be the subsets of \( X \) with cardinality \( n \). Clearly, if \( b \) is large with respect to \( n \), \( f(x) \approx 1/2n \) and so for any \( y \in Y \), \( \bar{f}(y) \approx 1/2 \), even for the \( y_* \) minimizing \( \bar{f}(y) \). At the other extreme, if \( b \) is small with respect to \( n \), \( \bar{f}(y_*) \) will be close to zero. For example, if \( b = n \), \( \bar{f}(y_*) = 0 \). It will follow from Theorem 8, that \( \bar{f}(y_*) \approx 0 \) for \( b \leq 2n \log 2 \). This model for generating a random probability gives insight into the following problem. If data is generated from a structureless model, random fluctuations may produce structure that is picked up by a rich enough data analytic procedure. As
b varies in the above model, the random probability converges to a uniform distribution. The following theorem gives an indication of how large b must be for all projections to be close to uniform. Some required notation: For \( \lambda < 0 \), let \( p_\lambda(j) = e^{-\lambda} \lambda^j/j! \) denote the Poisson density. Let \( P_\lambda(j) = \sum_{i=0}^{j} p_\lambda(i) \).

Let \( m \) be the largest integer with \( P_\lambda(m) \leq 1/2 \), \( P_\lambda(m+1) > 1/2 \). Define \( \theta = \theta(\lambda) \) by

\[
P_\lambda(m) + \theta P_\lambda(m+1) = \frac{1}{2}, \quad \text{so} \quad 0 \leq \theta < 1.
\]

When \( \lambda \) is an integer, Ramanujan showed that

\[
\theta = \frac{1}{3} + o(1/\lambda) \quad \text{as} \quad \lambda \to \infty.
\]

See Cheng (1949) for references and extensions of Ramanujan's results.

**Theorem 8.** Suppose that \( n \) and \( b \) tend to infinity in such a way that \( b/2n + \lambda \). Let \( y_* \) be the set with smallest value of \( \bar{f}(y_*) \). Then

\[
|\bar{f}(y_*) - \frac{1}{2}| + |\bar{f}(y_*^c) - \frac{1}{2}| = \frac{2e^{-\lambda} \lambda^m}{m!} (1 + \theta \times (\frac{\lambda}{m+1} - 1) + o_p(1)).
\]

**Remarks.** For \( \lambda \leq \log 2 \), \( m = 0 \) and the variation distance can be shown to tend to one. For large \( \lambda \), \( e^{-\lambda} \lambda^m/m! \approx 1/\sqrt{2\pi \lambda} \); thus for large \( \lambda \), the variation distance tends to zero like \( 1/\sqrt{\lambda} \). This is not very rapid as the following table shows (note that for integer \( \lambda \), \( m+1 = \lambda \), so the asymptotic value of the variation distance is \( 2e^{-\lambda} \lambda^m/m! \).

<table>
<thead>
<tr>
<th>( \lambda )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( e^{-\lambda} \lambda^m/m! )</td>
<td>.74</td>
<td>.54</td>
<td>.44</td>
<td>.40</td>
<td>.36</td>
<td>.32</td>
<td>.30</td>
<td>.28</td>
<td>.26</td>
<td>.24</td>
</tr>
</tbody>
</table>

**Proof of Theorem 8.** The argument will only be sketched. For \( b \) and \( n \) large, the number of balls in the ith box has a limiting Poisson distribution with parameter \( \lambda \), and different boxes can be treated as independent. The arguments in Diaconis and Freedman (1982) can be used to justify this step.
Thus let $X_1, X_2, \ldots,$ be independent Poisson variables with mean $\lambda$. With probability 1, eventually the median of $X_1, X_2, \ldots, X_{2n}$ is $m+1$ and the proportion of $X_i$ $1 \leq i \leq 2n$ equal to $j$ is $p_\lambda(j) + o(1)$ uniformly for $0 \leq j \leq m+1$. Let $S^-$ be the sum of the $n$ smallest $X_i$ $1 \leq i \leq 2n$. It follows that $S^-/2n$ equals

$$0 p_\lambda(0) + p_\lambda(1) + \ldots + m p_\lambda(m) + o(m+1) p_\lambda(m+1) + o(1).$$

This sum equals

$$\frac{\lambda}{2} - \frac{e^{-\lambda} \lambda^{m+1}}{m!} \left(1 + o(m+1) - 1\right) + o(1).$$

Thus $S^-/2\lambda n$, the limit of $\bar{F}(y_*)$, has the asserted limiting value. □
REFERENCES


