1

Symmetry, Classification, and the Analysis of Structured Data

1.1 Introduction

George Pólya, in his introduction to mathematics and plausible reasoning, observes that

> A great part of the naturalist's work is aimed at describing and classifying the objects that he observes. A good classification is important because it reduces the observable variety to relatively few clearly characterized and well-ordered types.

Pólya's (1954, p. 88) remark introduces us directly to the practical aspect of partitioning a large number of objects by exploring certain rules of equivalence among them. This is how symmetry will be understood in the present text: as a set of rules with which we may describe certain regularities among experimental objects or concepts. The classification of crystals, for example, is based on the presence of certain symmetries in their molecular framework, which in turn becomes observable by their optical activity and other measurable quantities.

The delicate notion of measuring something on these objects and recording their data is included in the naturalist's methods of description, so that the classification of the objects may imply the classification or partitioning of their corresponding data. Pólya's picture also includes the notion of interpreting, or characterizing, the resulting types of varieties. That is, the naturalist has a better result when he can explain why certain varieties fall into the same type or category.

This chapter is an introduction to the interplay among symmetry, classification, and experimental data, which is the driving motive underlying any symmetry study and is often present in the basic sciences. The purpose here is to demonstrate that principles derived from such interplay often lead to novel ways of looking at data, particularly of planning experiments and, potentially, of facilitating contextual explanation. We will observe the intertwined presence of symmetry, classification, and experimental data in a number of examples from chemistry, biology, and physics. Many principles and techniques will repeat across different disciplines, and it is exactly this cross-section of knowledge that constitutes the higher motivation and basis for these symmetry studies.

1.2 Symmetry and Classification

In grade school we were amused (for a little while at least!) by drawings and games with colorful patterns repeated periodically along straight lines and contours. These bands can be classified according to their distinct generating rules, such as horizontal translations, line and point reflections and rotations. These rules for symmetry in two dimensions are explored in wallpaper, textile, and tapestry designs, with the technical constraint of artistically and graphically designing these repeating motifs within the finite boundaries of the work.

The common understanding and perception of symmetry developed from our collective sensory and cultural experience with repetition or constancy can guide us in classifying, for example, the uppercase roman font printing of the English alphabet, imagined as subsets of the Euclidean plane. For example, the letters N, S, and Z are characterized by having a center of reflection symmetry whereas the letters H, I, O, and X have line (horizontal and vertical) and point reflection symmetry.

When a letter and its transformed image under a vertical line reflection $v : (y_1, y_2) \mapsto (-y_1, y_2)$ are indistinguishable, we say that the letter has the symmetry of v. If, in addition, the letter has the symmetry of a horizontal line reflection $h : (y_1, y_2) \mapsto (y_1, -y_2)$, then, consequently, it must have the symmetry of the iterated transformation (vh) of these two symmetries. Because the iterated transformation of h and v is a point reflection $o : y \mapsto -y$, we then learn that the letter has the symmetries of v, h, and o. Trivially, all letters have the symmetry of the identity transformation $1 : y \mapsto y$, often indicated simply as 1.

The resulting symmetries in $G = \{1, v, h, o\}$ *multiply* according to Table (1.1) and share the algebraic properties of a finite group: the product $(*)$ of two symmetries is a symmetry; the product is associative; 1 is the identity element and all symmetries have an inverse symmetry also in G.

$*$	1	v	h	o
1	1	v	h	o
v	v	1	o	h
h	h	o	1	v
o	o	h	v	1

(1.1)

We observe, in addition, that any $f \in G$ is a bijective transformation of the Euclidean plane preserving its algebraic properties, in the sense that $f(x + y) = f(x) + f(y)$ for all vectors x, y in the plane. These are called automorphisms of the plane.

Any two letters are then classified together when they share the same set of symmetries or automorphisms. For example, the letters $\ell \in \{\text{H,I,O,X}\}$ are classified together by sharing the symmetries of G. We then say that G is their automorphism group and write $\text{Aut}\{\ell\} = G$ for all $\ell \in \{\text{H,I,O,X}\}$. In summary, after classifying the letters of the English alphabet, we have the following:

ℓ	$\text{Aut}(\ell)$
F,G,J,K,L,P,Q,R	1
A,M,T,U,V,W,Y	$1, v$
B,C,D,E	$1, h$
N,S,Z	$1, o$
H,I,O,X	$1, h, v, o$

1.3 Data Indexed by Symmetries

The lines in the left-hand side of Table (1.2) were abstracted from a visual acuity testing chart developed for the Early Treatment Diabetic Retinopathy Study, or ETDRS (Ferris III et al., 1993, Table 5). The 10 different letters {Z,N,H,V,R,K,D,S,O,C} that appear in the actual chart differ only in that they are printed with specially created Sloan fonts (Sloan, 1959) and are presented according to an experimental protocol.

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

ℓ	$\text{Aut}(\ell)$	$p(\ell)$	entropy(ℓ)	$-\log \text{CS}(\ell)$
Z	$1, o$	0.844	0.433	0.63
N	$1, o$	0.774	0.535	0.53
H	$1, o, v, h$	0.688	0.619	0.44
V	$1, v$	0.636	0.656	0.56
R	1	0.622	0.663	0.46
K	1	0.609	0.669	0.57
D	$1, h$	0.556	0.687	0.43
S	$1, o$	0.516	0.693	0.44
O	$1, o, v, h$	0.470	0.692	0.34
C	$1, h$	0.393	0.673	0.36

(1.2)

The individual letters are shown in the adjacent table, along with their automorphisms, estimated probability (p) of correct identification, corresponding entropy $-[p \log p + (1-p)\log(1-p)]$, and estimated ($-\log$) contrast sensitivity. The entropy of a letter is a measure of the relative uncertainty in its correct identification. Its value is zero in the absence of uncertainty, and it is positive otherwise and attains its maximum value ($\log 2 = 0.693$) when the events are equally like, that is, $p = 1/2$. The probabilities of correct identification were estimated from a large sample of test subjects reported by Ferris III et al. (1993). The letter contrast sensitivity is a direct measure of the subject's visual performance. It is estimated from psychophysical experiments to determine the threshold of perception under varying levels of background contrast (Alexander et al., 1997). The smaller is the contrast needed to see the letter, the larger is the sensitivity.

We are interested in describing the connection among font symmetry, letter entropy, and contrast sensitivity from samples of Sloan lines similar to those shown in (1.2).

To each symmetry t in $G = \{1, v, h, o\}$, indicate by fix_t the subset of letters in a selected line with the symmetry of t and by $x_t = |\text{fix}_t|$ the number of elements in fix_t. For example, the first line C O H Z V in the chart gives

$$(1, v, h, o) \xrightarrow{x} (5, 3, 3, 3), \tag{1.3}$$

which is an example of data indexed by the elements in G, and a point in the vector space $\mathcal{V} = \mathbb{R}^4$. If $|\text{fix}_t| \neq 0$ then the mean line entropy

$$\frac{1}{|\text{fix}_t|} \sum_{\ell \in \text{fix}_t} \text{entropy}(\ell)$$

based on those letters with the symmetry of t leads to a different indexing of data by the elements of G. In this case, for the same line, the new indexing is

$$(1, v, h, o) \xrightarrow{x} (0.512, 0.655, 0.661, 0.575). \tag{1.4}$$

Similarly, when averaging the ($-$ log) contrast sensitivity over the letters with same symmetry, the indexing is

$$(1, v, h, o) \xrightarrow{x} (0.466, 0.446, 0.380, 0.476). \tag{1.5}$$

Note that the first components in (1.3), (1.4), and (1.5) are, respectively, the total number (5) of letters in each line, the line mean entropy and mean contrast sensitivity. These are examples of data indexed by a particular structure (a finite group in this case) or, simply, examples of structured data.

If similar lines are sampled from a larger set of charts, then x is a random vector and statistical summaries of the resulting sample are of interest. For example, Figure 1.1 summarizes the distributions of the four entropy components in (1.4) based on a sample of 42 lines similar to those in (1.2). The distributions should be interpreted along with the symmetry content of the underlying set of Sloan letters and the likely distribution of these symmetries over the 42 lines. Table (1.6) summarizes the underlying joint distribution of the 10 reference letters and symmetries. The marginal column and row sums are, respectively, the number

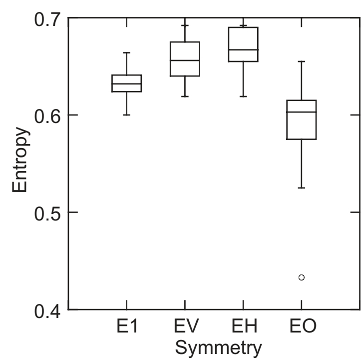

Figure 1.1: Distribution of line mean letter entropy by symmetry type.

$|\mathrm{Aut}(\ell)|$ of automorphisms of ℓ and the number $|\mathrm{fix}_t|$ of letters with the symmetry of t.

$t\backslash\ell$	Z	N	H	V	R	K	D	S	O	C	$\lvert\mathrm{fix}_t\rvert$
1	1	1	1	1	1	1	1	1	1	1	10
v	0	0	1	1	0	0	0	0	1	0	3
h	0	0	1	0	0	0	1	0	1	1	4
o	1	1	1	0	0	0	0	1	1	0	5
$\lvert\mathrm{Aut}(\ell)\rvert$	2	2	4	2	1	1	2	2	4	2	22

(1.6)

It is observed that point symmetry is present in the largest number ($|\mathrm{fix}_t| = 5$) of reference letters and that at the same time the two letters with the smallest entropy (Z and N) have $|\mathrm{Aut}(\ell)| = 2$ characterized precisely by the same symmetry.

1.4 Symmetry and Data Reduction

Classical physical measurements are understood, mathematically, as real vectors x in the usual Euclidean vector space. Consequently, it is of natural interest to represent the symmetries described by $G = \{1, h, v, o\}$ into the vector space $\mathcal{V} = \mathbb{R}^4$ for the data, shown in (1.3), (1.4), or (1.5), indexed by G. These representations are accomplished by associating to each element t in G a linear transformations ρ_t in $\mathcal{V}$.

Specifically, using the multiplication table of G shown in (1.1), to each element t in G associate the permutation matrix

$$\{e_1, e_v, e_h, e_o\} \xrightarrow{\rho_t} \{e_{t*1}, e_{t*v}, e_{t*h}, e_{t*o}\}, \tag{1.7}$$

in which the entry $(\rho_t)_{sf}$ of ρ_t at row s and column f is equal to 1 if and only if $f = t * s$, for $f, t, s \in G$. For example, $(\rho_v)_{ho} = 1$ indicates that $v * h = o$. Therefore,

$$\rho_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \rho_v = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix},$$

$$\rho_h = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}, \quad \rho_o = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

These resulting linear transformations then connect the symmetries in the group G with the vector space $\mathcal{V}$ for (1.3), (1.4), or (1.5) in a way that the multiplication in G described by (1.1) is now represented as multiplication of nonsingular linear transformations in $\mathcal{V}$, that is,

$$\rho_{t*t'} = \rho_t \rho_{t'} \quad \text{for all } t, t' \in G. \tag{1.8}$$

This is the homomorphic property, characteristic of these linear representations.

The algebraic aspects developed in the next chapters will show that certain linear combinations of $\{\rho_1, \rho_v, \rho_h, \rho_o\}$ then lead to four algebraically orthogonal projection matrices $\mathcal{P}_1, \ldots, \mathcal{P}_4$, given by

$$1/4 \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \; 1/4 \begin{bmatrix} 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ -1 & -1 & 1 & 1 \\ -1 & -1 & 1 & 1 \end{bmatrix},$$

$$1/4 \begin{bmatrix} 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 1 & -1 & 1 \end{bmatrix}, \; 1/4 \begin{bmatrix} 1 & -1 & -1 & 1 \\ -1 & 1 & 1 & -1 \\ -1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}, \tag{1.9}$$

respectively, which determine statistical summaries $\mathcal{P}_1 x, \ldots, \mathcal{P}_4 x$ characterized by the particular representation (1.7) of G. We will refer to these summaries, in general, as the canonical invariants in the study – a concept that will be developed throughout the text. In the present case, these projections directly identify four invariants, namely,

$$\mathcal{I}_1 = x_1 + x_o + x_v + x_h, \quad \mathcal{I}_v = x_1 + x_v - x_o - x_h,$$
$$\mathcal{I}_h = x_1 + x_h - x_o - x_v, \quad \mathcal{I}_o = x_1 + x_o - x_v - x_h, \tag{1.10}$$

Cambridge University Press
978-0-521-84103-0 — Symmetry Studies: An Introduction to the Analysis of Structured Data in Applications
Marlos A. G. Viana
Excerpt
More Information

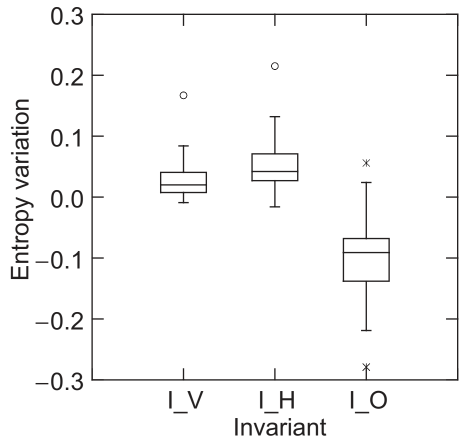

Figure 1.2: Distribution of the canonical invariants $\mathcal{I}_v, \mathcal{I}_h, \mathcal{I}_o$ for the mean line entropy data.

each one taking values on subspaces in the dimension of 1. These summaries depend on the labels (provided by G) only up to companion points determining a linear subspace of the data space (called invariant subspace). For example, the summary $x_1 + x_o - x_v - x_h$ is such that

$$x_{t*1} + x_{t*o} - x_{t*v} - x_{t*h} = \pm(x_1 + x_o - x_v - x_h) \quad \text{for all} \quad t \in G.$$

The summaries of the data induced by G can then be interpreted as of exactly two types:

(1) The overall sum of responses ($\mathcal{I}_1$) and
(2) The three pairwise comparisons ($\pm\mathcal{I}_v$, $\pm\mathcal{I}_h$, $\pm\mathcal{I}_o$).

These pairwise comparisons are the basis for inferences in this particular symmetry study. Figure 1.2 summarizes the distributions of the canonical invariants $\mathcal{I}_v, \mathcal{I}_h, \mathcal{I}_o$ based on 42 lines of Sloan fonts.

The invariants are the data that should be retained when the arbitrariness of where is left (right) and where is up (down), associated with the action (1.7), is resolved. For example, then, $x_1 + x_o - x_v - x_h$ compares point and line symmetries in a way that depends on the chosen planar orientation only up to an invariant subspace. As effectively suggested by Weyl (1952, p. 144),

> Whenever you have to do with a structure-endowed entity try to determine its group of automorphisms, the group of those element-wise transformations which leave all structural relations undisturbed. You can expect to gain a deep insight into its constitution this way.

We observe that the derivation of these data summaries depends only on the set of labels and the symmetries of interest. Any subsequent statistical analysis, of course, would include the assumptions that apply to a particular experimental condition. For example, if the data indexed by G are the frequency distributions

$$x_1 = (0, 42), \quad x_o = (21, 21), \quad x_v = (39, 3), \quad x_h = (32, 10)$$

with which the corresponding symmetries appeared in at most 2 or in 3 or more of the 5 letters in each line, respectively, summed over 42 Sloan lines, then the invariants may be interpreted as three pairwise comparisons

$$\begin{aligned} x_1 + x_o &= (21, 63) \quad \text{vs.} \quad x_v + x_h = (71, 13), \\ x_1 + x_h &= (32, 52) \quad \text{vs.} \quad x_v + x_o = (60, 24), \\ x_1 + x_v &= (39, 45) \quad \text{vs.} \quad x_o + x_h = (53, 31) \end{aligned} \tag{1.11}$$

between these frequency distributions, which, statistically, could be carried out in many different ways.

1.5 Statistical Aspects

We have remarked that the matrices $\mathcal{P}$ in (1.9) lead to the data summaries $\mathcal{P}x$ shown in (1.10). These matrices are algebraically orthogonal ($\mathcal{P}_i\mathcal{P}_j = \mathcal{P}_j\mathcal{P}_i = 0$ for $i \neq j$) projections ($\mathcal{P}_i^2 = \mathcal{P}_i$, $i = 1, \ldots, 4$) that reduce the identity operator I in the data vector according to the sum

$$I = \mathcal{P}_1 + \mathcal{P}_2 + \mathcal{P}_3 + \mathcal{P}_4,$$

so that, consequently, the theory of statistical inference for (real symmetric) quadratic forms can be applied to study the decomposition

$$x'x = x'\mathcal{P}_1x + \cdots + x'\mathcal{P}_4x$$

of the sum of squares $x'x$ of x.

To illustrate, consider the data shown in (1.12). Each row is a sample of size 5, obtained from 5 different Sloan chart lines, of the corresponding mean line entropy $x_t = \sum_{\ell \in \text{fix}_t} \text{entropy}(\ell)/|\text{fix}_t|$, indexed by the symmetry element t.

t\Sample	1	2	3	4	5
1	0.614	0.636	0.632	0.624	0.66
v	0.675	0.619	0.692	0.640	0.619
h	0.655	0.619	0.660	0.690	0.667
o	0.603	0.603	0.553	0.603	0.635

(1.12)

The application of the algebraic arguments outlined above and detailed in the next chapters resulted in the analysis of variance table shown in (1.13), where the degrees of freedom (df) are the traces of the corresponding canonical projections and the F-ratios derived from the ratios of the mean sum of squares $x'\mathcal{P}x/df$ relative to the mean error sum of squares.

Component	$x'\mathcal{P}x$	df	$x'\mathcal{P}x/df$	F-ratio
$\mathcal{I}_1$	8.0633	1	8.0633	
$\mathcal{I}_v$	0.000757	1	0.000757	1.036
$\mathcal{I}_h$	0.002312	1	0.002312	3.165
$\mathcal{I}_o$	0.006956	1	0.006956	9.525
Error	0.011684	16	0.000730	

(1.13)

Here, the decomposition of the sum of squares is the consequence of jointly shuffling the rows and columns of the table in (1.12) using $G = \{1, h, v, o\}$ and the permutations of $\{1, 2, 3, 4, 5\}$, respectively.

Shuffling the rows in (1.12) according to G means relabeling them according to

$$\begin{bmatrix} v \\ 1 \\ o \\ h \end{bmatrix}, \quad \begin{bmatrix} h \\ o \\ 1 \\ v \end{bmatrix}, \quad \text{or} \quad \begin{bmatrix} o \\ h \\ v \\ 1 \end{bmatrix},$$

the result of multiplying the original first column by v, h, o respectively. On the other hand, shuffling the columns, indexed by $\{1, 2, 3, 4, 5\}$, simply means performing all their permutations.

Under the usual normality assumptions and corresponding hypotheses of the form $\mathcal{I} = 0$ (in terms of expected values), the indicated F-ratios have a central F-distribution with degrees of freedom 1 and 16 and can be used to test these parametric hypotheses.

It is now evident that the same canonical invariants $\mathcal{I} = \mathcal{P}x$ can be the object of descriptive summaries (Figure 1.2), nonparametric comparisons (1.11), or parametric hypotheses (1.13) for the structured data.

The analysis of variance (1.13) points to a significant distinction in mean line entropy when the differentiation (among chart lines) is due to point vs. line symmetries ($\mathcal{I}_o \neq 0$). The explanation of this finding, expressed in terms of the invariant $\mathcal{I}_o$, may then be found in the theories of eye movement, for example.

1.6 Algebraic Aspects

The role of algebra in the analysis of structured data is that of ascertaining its methodological aspects, of providing a well-defined sequence of steps leading to predictable data-analytic tools. We illustrate this with the following preliminary summary.

The mean line entropy data $x' = (x_1, x_v, x_h, x_o)$ shown in Table (1.12) were introduced as an example of data indexed by the elements of a finite group $G = \{1, v, h, o\}$. It was then possible to identify

(1) a set (G) of labels with the algebraic properties of a finite group;
(2) a set of data (x) indexed by those labels (the structured data);
(3) a group action, defined in (1.7), with which the symmetries in G were applied to itself;
(4) a linear representation (ρ) of that action connecting the labels and the data vector space ($\mathcal{V}$);
(5) the projection matrices $\mathcal{P}_1, \ldots, \mathcal{P}_4$ shown in (1.9);
(6) the canonical invariants $\mathcal{P}_1 x, \ldots, \mathcal{P}_4 x$ in the data, described in (1.10), and their interpretations, and
(7) the resulting analysis of variance $x'x = x'\mathcal{P}_1 x + \cdots + x'\mathcal{P}_4 x$ based on the decomposition $I = \mathcal{P}_1 + \cdots + \mathcal{P}_4$, shown in (1.13) .

Note that the effect of reordering the basis used in the construction (1.7) of the representation ρ is such that the new decomposition is now

$$I = \eta \mathcal{P}_1 \eta' + \cdots + \eta \mathcal{P}_4 \eta'$$

where η is the corresponding permutation matrix. The new decomposition is in fact the same as (1.9), but relabeled. For example, if the entries had been written in the order of $1, o, v, h$ instead of the original order $1, v, h, o$, then $\eta \mathcal{P}_4 \eta' = \mathcal{P}_3$, $\eta \mathcal{P}_3 \eta' = \mathcal{P}_2$, and $\eta \mathcal{P}_2 \eta' = \mathcal{P}_4$. Consequently, the invariants (1.10), their interpretation, and the resulting analysis of variance (1.13) would remain exactly the same.

However, the algebra has more to say here. A quick review of the projection matrices in (1.9) reveals that they can be written in terms of the matrices

$$\mathcal{A} = \frac{1}{2}\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad \mathcal{Q} = \frac{1}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \tag{1.14}$$

which combine and compare the two components of a point in $\mathbb{R}^2$ and orthogonally reduce, or decompose, the identity matrix in that space into the sum $\mathcal{A} + \mathcal{Q}$. This reduction in $\mathbb{R}^2$ is an example of a *standard reduction* and will be used many times in these studies.

We have, using the symbol $\otimes$ to indicate the Kronecker product of two matrices, that

$$\mathcal{P}_1 = \mathcal{A} \otimes \mathcal{A}, \quad \mathcal{P}_2 = \mathcal{Q} \otimes \mathcal{A}, \quad \mathcal{P}_3 = \mathcal{A} \otimes \mathcal{Q}, \quad \mathcal{P}_4 = \mathcal{Q} \otimes \mathcal{Q}.$$

If, in addition, the data x can justifiably be indexed by a product $f \otimes g$ of two two-level labels f and g, then the data (briefly identified here with the labels) decompose as $\mathcal{A}f \otimes \mathcal{A}g$, $\mathcal{Q}f \otimes \mathcal{A}g$, $\mathcal{A}f \otimes \mathcal{Q}g$, and $\mathcal{Q}f \otimes \mathcal{Q}g$. This, more elementary, construction of the projections $\mathcal{P}_1, \ldots, \mathcal{P}_4$ is explained in terms of smaller component symmetry groups acting (by simple transpositions) on the bivariate component labels f, g. It leads, precisely, to the well-known concepts of factors and factor levels in simple factorial experiments. It is only when these component groups are introduced that a distinction between the projections $\{\mathcal{P}_2, \mathcal{P}_3\}$ and $\mathcal{P}_4$ can be envisioned.