Chapter 8. Spectral Analysis

Often data are presented as a function $f(x)$ defined on some index set X. If X is connected to a group, the function f can be "Fourier expanded" and one may try to interpret its coefficients. This is a rich idea which includes the usual spectral analysis of time series and the analysis of variance.

This chapter develops the idea in stages. First, for data on groups (time series, binary vectors, and permutation data). Then the idea is developed for data on homogeneous spaces (the sphere and partially ranked data). Next some theory needed for practical computation is derived. All of this is illustrated for some classical designs in the analysis of variance. Finally, some research projects are spelled out.

A. Data on groups

1. *Time series analysis.* Fourier analysis of time series or other signals is a familiar scientific activity. For example, data on the number of babies born daily in New York City over a five year period are studied by Izenman and Zabell (1978). Here $X = \{1, 2, \ldots, n\}$ with $n = 5 \times 365 + 1$. The data are represented as a function $f(x) = \#$ born on day x. Inspection of this data shows strong periodic phenomena: about 450 babies are born on each week day and about 350 on each day of the weekend. (Physicians don't like to work on weekends.) There might also be monthly or quarterly effects.

To examine (and discover) such phenomena, scientists pass from the original data $f(x)$ to its Fourier transform

$$\hat{f}(y) = \Sigma_x f(x)\, e^{2\pi i x y/n}$$

where the sum runs over $x = 0, 1, \ldots, n-1$. Fourier inversion gives

$$f(x) = \frac{1}{n} \Sigma_y \hat{f}(y)\, e^{-2\pi i x y/n}.$$

It sometimes happens that a few values of $\hat{f}(y)$ are much larger than the rest and determine f in the sense that f is closely approximated by the function defined by using only the large Fourier coefficients in the inversion formula. When this happens, we have f approximated by a few simple periodic functions of x, e.g. $e^{-2\pi i x y/n}$, and may feel we understand the situation.

The hunting and interpretation of periodicities is one use of spectral analysis. A splendid introduction to this subject is given by Bloomfield (1976). A more advanced treatment from the same point of view is given by Brillinger (1975).

There are other interpretations of spectral analysis. The discussion papers by Jenkins (1961) and Parzen (1961) present a useful survey of different views. Some further discussion is given in the last section of this chapter. For now, we will stick to the data analytic view outlined above.

2. *Permutation data.* Spectral analysis can be carried out for any group using the matrix entries of the irreducible representations as a basis. Before developing this in general, here is an example.

In Chapter 5-A we discussed rankings of three items. People were asked to rank where they wanted to live: in a city, suburbs, or country. The rankings were

π	city	suburbs	country	#
id	1	2	3	242
(23)	1	3	2	28
(12)	2	1	3	170
(132)	3	1	2	628
(123)	2	3	1	12
(13)	3	2	1	359

Here $X = S_3$, and $f(\pi)$ is the number choosing π. There are three irreducible representations of S_3: the trivial, sgn, and two-dimensional representation ρ. The Fourier inversion theorem gives

$$f(\pi) = \frac{1}{6}\{\hat{f}(\text{triv}) + \text{sgn}(\pi) \cdot \hat{f}(\text{sgn}) + 2Tr(\rho(\pi^{-1})\hat{f}(\rho)).$$

Expanding the trace gives a spectral analysis of f as a sum of orthogonal functions.

To facilitate comparison between functions in this basis, let us choose an orthogonal version of ρ. Thus, using cycle notation on S_3

$$\begin{array}{cccccccc} \pi & id & (1\,2) & (2\,3) & (1\,3) & (1\,2\,3) & (1\,3\,2) \\ \rho(\pi) & \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} & \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} & \frac{1}{2}\begin{pmatrix} 1 & \sqrt{3} \\ \sqrt{3} & -1 \end{pmatrix} & \frac{1}{2}\begin{pmatrix} 1 & -\sqrt{3} \\ -\sqrt{3} & -1 \end{pmatrix} & \frac{1}{2}\begin{pmatrix} -1 & -\sqrt{3} \\ \sqrt{3} & -1 \end{pmatrix} & \frac{1}{2}\begin{pmatrix} -1 & \sqrt{3} \\ -\sqrt{3} & -1 \end{pmatrix} \end{array}$$

These are arrived at by choosing $w_1 = \frac{1}{\sqrt{2}}(e_1 - e_2)$, $w_2 = \frac{1}{\sqrt{6}}(e_1 + e_2 - 2e_3)$ as an orthonormal basis for $\{v \in \mathbb{R}^3 : v_1 + v_2 + v_3 = 0\}$. The matrices $\rho(\pi)$ give the action of π in this basis. Now

$$\hat{f}(\text{triv}) = 1439,\ \hat{f}(\text{sgn}) = 242 - 28 - 170 + 628 + 12 - 359 = 325,$$
$$\hat{f}(\rho) = \begin{pmatrix} -54.5 & 285\sqrt{3}/2 \\ -947\sqrt{3}/2 & -101.5 \end{pmatrix}.$$

Define four functions on S_3 by

$$\sqrt{2}\ \rho(\pi^{-1}) = \begin{pmatrix} a(\pi) & b(\pi) \\ c(\pi) & d(\pi) \end{pmatrix}.$$

With this definition, the functions 1, sgn π, $a(\pi)$, $b(\pi)$, $c(\pi)$, $d(\pi)$ are orthogonal and have the same length.

Expanding the trace in the Fourier inversion theorem gives

$$f(\pi) = \frac{1}{6}\{1439 + 325\ \mathrm{sgn}(\pi) - 54.5\sqrt{2}\ a(\pi) - 947\sqrt{3/2}\ b(\pi) + 285\sqrt{3/2}\ c(\pi) - 101.5\sqrt{2}\ d(\pi)\}.$$

$$\doteq \frac{1}{6}\{1439 + 325\ \mathrm{sgn}(\pi) - 77\ a(\pi) - 1160\ b(\pi) + 349\ c(\pi) - 144\ d(\pi)\}.$$

As a check, when $\pi = \mathrm{id}$, this becomes

$$242 = \frac{1}{6}\{1439 + 325 - 109 - 203\}.$$

The largest non-constant coefficient, 1160, multiplies $b(\pi)$. This is the function

π	id	(1 2)	(2 3)	(1 3)	(1 2 3)	(1 3 2)
$b(\pi)$	0	0	$\sqrt{3/2}$	$-\sqrt{3/2}$	$\sqrt{3/2}$	$-\sqrt{3/2}$

or

$$b(\pi) = \begin{cases} -\sqrt{3/2} & \text{if cities are ranked 3rd } (\pi(1) = 3) \\ 0 & \text{if country is ranked 3rd } (\pi(3) = 3) \\ \sqrt{3/2} & \text{if suburbs are ranked 3rd } (\pi(2) = 3). \end{cases}$$

Spectral analysis gives fresh insight into this little data set: After the constant, the best single predictor of f is what people rank last. Now $b(\pi)$ enters with a *negative* coefficient. It "follows" that people "hate" the city most, the suburbs least and the country in between. Going back to the data $\#\{\pi(1) = 3\} = 981$, $\#\{\pi(2) = 3\} = 40$, $\#\{\pi(3) = 3\} = 412$, so the effect is real. It seems at variance with the unfolding hypothesis, but is in fact necessary. If people make individual rankings by unfolding, the proportions connected to extreme ranks will be monotone with the unfolding point in the center.

The kind of data analytic interpretation given in the last paragraph seems mandatory – we must seek subject matter interpretation of our findings. As a word of warning, almost any set of summary statistics can have a story woven about them – we are good at making up stories.

EXERCISE 1. Carry out the analysis of this section for the non-orthogonal representation of S_3 given in Chapter 2-A. Show that the main conclusions do not change.

3. *A general case, with inference.* For G a finite group, and f a function on G, the Fourier inversion theorem gives

$$f(s) = \frac{1}{|G|}\Sigma_\rho\ d_\rho\ Tr(\rho(s^{-1})\hat{f}(\rho)).$$

Now let ρ be a unitary representation, so $\rho(t)^* = \rho(t^{-1})$. Corollaries 2 and 3 to Schurs lemma of Chapter 2 yield orthogonality relations for the matrix entries of ρ. Define the usual inner product on all functions on G: $< \phi|\psi > = \frac{1}{|G|}\Sigma\, \phi(t)\psi(t)^*$. Then

$< \rho_{ij}|\eta_{k\ell} > = 0$ if ρ and η are inequivalent unitary representations, for any $ij, k\ell$.

$$< \rho_{ij}|\rho_{k\ell} > = \begin{cases} 0 & \text{unless } i = k \text{ and } j = \ell, \\ \frac{1}{d_\rho} & \text{if } i = k,\ j = \ell. \end{cases}$$

It follows that the functions $\tilde{\rho}_{ij}(s) = \sqrt{d_\rho}\ \rho_{ij}(s^{-1})$ are orthonormal on G with respect to $< \cdot|\cdot >$.

To numerically compute the spectral representation, compute $\hat{f}(\rho)$ and expand the trace giving

$$f(s) = \frac{1}{|G|} \sum_{\rho,i,j} \sqrt{d_\rho}\ \hat{f}(\rho)_{ji}\ \tilde{\rho}_{ij}(s).$$

The squared length of the projection indexed by functions associated to ρ is $d_\rho\ Tr(\hat{f}(\rho)\hat{f}(\rho)^*) = d_\rho \|\hat{f}(\rho)\|^2$.

Elementary inference. In carrying out spectral analysis, it is natural to wonder "if the data had come out a bit different, would the inferences change?" This is susceptible to a wealth of interpretations – from an elementary sensitivity analysis, through a bootstrap analysis, through a frequentist model through a Bayesian analysis. At present, very little is available in off the shelf tools. We here develop the obvious normal theory. Some more speculative suggestions are contained in the final section of this chapter.

Let G be a finite group. Suppose that an observed function $f(s)$ can be regarded as a true function $\mu(s)$ plus an error or perturbation function $\varepsilon(s)$

$$f(s) = \mu(s) + \varepsilon(s).$$

The strongest possible assumptions that can be made are $\mu(s)$ fixed (or known) and $\varepsilon(s) \sim N(0, \sigma^2)$ (normal, with mean 0 and variance σ^2 independent for each s). Then for an orthogonal representation ρ, the coefficients $\hat{f}(\rho)_{ij}\sqrt{d_\rho}$ are all independent normals, with mean $\sqrt{d_\rho}\hat{\mu}(\rho)_{ij}$ and variance $\sigma^2|G|$. (I have assumed that all of the representations are real. For unitary representations, complex normal distributions occur). Further, for ρ and η inequivalent representations, $\hat{f}(\rho)_{ij}$ and $\hat{f}(\eta)_{k\ell}$ are independent.

EXERCISE 2. Prove the results in the last paragraph.

If σ^2 is assumed known, or $\hat{\mu}(\rho) = 0$ is assumed for some irreducible so σ^2 can be estimated, all of the usual inferential tools associated to the general linear model are available. In particular, $d_\rho\|\hat{f}(\rho)\|^2$ is distributed as σ^2 times

a chi-squared variable on d_ρ^2 degrees of freedom (if $\hat{\mu}(\rho) = 0$). Fisher's test for the significance of the largest Fourier coefficient (see e.g. Bloomfield (1976) or Anderson (1971)) is easily adapted to this setting.

If σ^2 is assumed to depend on s, a rich array of tools from variance components and classical multivariate analysis are available. Work of the Danish school is particularly relevant – see Perlman (1988) for entry into this literature.

The normality of the Fourier transforms is approximately true under much less restrictive assumptions than normality of $f(s)$. After all, $\hat{f}(\rho)_{ij}$ is an average of a lot of things, and will be approximately normal and independent of other coefficients under fairly mild assumptions. This deserves to be worked out more carefully. Anderson (1971, Sec. 2.6), Freedman and Lane (1980), or Diaconis and Freedman (1984) are useful places to start.

This seems like a good place to point out

Elementary fact. Let ρ be an irreducible unitary representation of G. Let $L(G)$ be all functions on G. Consider the subspace L_j spanned by the matrix entries of the jth column of ρ. Then L_j is an invariant subspace isomorphic to ρ.

Proof. Let $s\ \rho_{ij}(t) = \rho_{ij}(s^{-1}t)$. The matrix equation $\rho(s^{-1})\rho(t) = \rho(s^{-1}t)$ shows that $\rho_{ij}(s^{-1}t) = \Sigma_k\ a_k \rho_{kj}(t)$ as a function of t. Thus L_j is an invariant subspace. Choosing the functions ρ_{ij}, $1 \le i \le d_\rho$ as a basis gives $\rho(s)$ being the associated matrix representation. □

Remarks. As j varies, the representations L_j give the d_ρ copies of ρ that appear in the decomposition of the regular representation. In applications, the columns of ρ sometimes have a natural meaning. For example, on S_n the jth column of the $n-1$-dimensional representation codes which i is mapped to position j in one choice of basis. Expanding the trace preserves this interpretation. Further discussion of interpretability of the basis functions appears in the last section of this chapter.

4. *Bahadur's item analysis.* Bahadur (1961) introduced a spectral analysis when $X = Z_2^k$ consists of data on binary k-tuples. For example, fix k at 5 and consider a population of families with five children. Take $f(x)$ as the proportion of families whose children were born in birth order x – so if $x = 01010$, the birth order was girl, boy, girl, boy, girl. Klotz (1970) gives an analysis of such data.

Economics has massive amounts of "panel study data." Here k might be 12 and x might represent the pattern of employed/unemployed behavior for a person in the study. It is not uncommon to have samples larger than 5,000 people. Hsiao (1986) is a recent book on this subject.

Bahadur's original motivation was test score analysis, where x records the pattern of correct/incorrect guesses, and $f(x)$ records the proportion of students answering in a given pattern.

Let x_i, $1 \le i \le k$ be the coordinate projection from Z_2^k. Let $f(x)$ be a probability on Z_2^k. Define $\alpha_i = E_f(x_i)$ and $z_i = (x_i - \alpha_i)/\sqrt{\alpha_i(1-\alpha_i)}$. Define

$$
\begin{aligned}
r_{ij} &= E_f(z_i z_j),\ i < j \\
r_{ij\ell} &= E_f(z_i z_j z_\ell),\ i < j < \ell \\
&\vdots \\
r_{1\ldots k} &= E_f(z_1 z_2 \ldots z_k).
\end{aligned}
$$

The $\binom{k}{2}$ parameters r_{ij} are correlations. The $r_{ij\ell}$ can be thought of higher order correlations, etc. Bahadur wrote the spectral decomposition of f as follows.

PROPOSITION 1. *(Bahadur). For f a probability on Z_2^k,*

$$f = f_1 \cdot f_2$$

where

$$f_1(x) = \Pi\ \alpha_i^{x_i}(1-\alpha_i)^{1-x_i} \text{ is the product distribution with margins matching } f.$$

$$f_2(x) = 1 + \sum_{i<j} r_{ij} z_i z_j + \sum_{i<j<\ell} r_{ij\ell} z_i z_j z_\ell + \ldots + r_{1\ldots k} z_1 \ldots z_k.$$

Proof. Consider the vector space of all functions on Z_2^k with inner product $< g, h >= E_{f_1}(g \cdot h)$. The set

$$S = \{1; z_1, z_2, \ldots, z_k;\ z_1 z_2, \ldots, z_{k-1} z_k;\ z_1 z_2 z_3, \ldots; \ldots z_1 \ldots z_k\}$$

consists of orthonormal functions: $\|g\| = 1$ for $g \in S$ and $< g, h >= 0$ for $g, h \in S$ but $g \neq h$. There are 2^k functions in S, so they form a basis. Now, set $f_2 = f/f_1$. Then $< f_2, g >= \Sigma f_2 g f_1 = E_f(g)$. Also $E_f(1) = 1, E_f(z_i) = 0$, so the proposition follows. □

The function f_2 measures departure from independence. It is natural to look at the individual coefficients to try to understand the nature of the dependence. Let $\delta_{(k)}^2 = \|f_2 - 1\|^2$. Thus define

$$
\begin{aligned}
\delta_{(k)}^2 &= \sum_{i<j} r_{ij}^2 + \sum_{i<j<\ell} r_{ij\ell}^2 + \ldots + r_{1\ldots k}^2 \\
&\stackrel{\mathrm{d}}{=} \delta_2^2 + \ldots + \delta_k^2.
\end{aligned}
$$

The ratios $\delta_j^2/\delta_{(k)}^2$ provide an index of the relative importance of the jth order terms. Similarly, if f_m is the approximation to f in which all terms in f_2 involving a product of more than m z_i's are omitted, then

$$\|f_m/f_1\|^2/\|f/f_1\|^2 = \frac{1 + \delta_1^2 + \ldots + \delta_m^2}{1 + \delta_1^2 + \ldots + \delta_k}$$

provides an index of quality for the approximation of f by f_m.

The analysis above is just the standard spectral analysis shifted to a more natural base measure (matched marginals versus uniform). As Bahadur remarks, it is also natural to expand and analyze $\log f$. This gets away from the problems of interpreting negative coefficients and links into the usual log-linear analysis of such data as a $2 \times 2 \times \ldots \times 2$ contingency table.

B. Data on homogeneous space

A natural and useful extension of spectral analysis occurs for homogeneous spaces. These were defined and illustrated in Chapter 3F.

1. *Definitions.* Let X be a finite set. Let G act transitively on X with N the isotropy subgroup: $N = \{s \in G: sx_0 = x_0\}$ where x_0 is some fixed point in X. Let $L(X)$ be all functions from X into the complex numbers C. Then $L(X)$ decomposes into irreducibles as

$$V_0 \oplus V_1 \oplus \ldots \oplus V_k.$$

Suppose we are given $f(x) \in L(X)$, regarded as data – $f(x)$ is the number (or proportion) of people having property x.

Spectral analysis is the decomposition of $f(x)$ into its projections on the irreducible invariant subspaces of $L(X)$, and the approximation of $f(x)$ by as small a number of projections as give a reasonable fit.

Here $L(X)$ is regarded as an inner product space using $< f|g > = \frac{1}{|X|}\Sigma f(x)g(x)^*$, and projections are orthogonal. Of course we usually want even more: the coefficients of f projected into V_i in some natural or interpretable basis help connect the analysis to the original subject matter.

As shown in Corollary 1 of Section 3 below the decomposition of the regular representation falls into this domain. The following are less standard.

Example 2. Partially ranked data. Let λ be a partition of n. Consider data consisting of partial rankings of n items of shape λ: thus people rank their favorite λ_1 items (but not within) and their next λ_2 items (but not within) and their final λ_k items (but not within). Here $n = \lambda_1 + \lambda_2 + \ldots + \lambda_k$ and we do not assume λ_i are ordered. Chapter 5B gives examples.

If $S_{\lambda_1} \times S_{\lambda_2} \times \ldots \times S_{\lambda_k}$ denotes the subgroup of S_n which allows permutations among the first λ_1 coordinates, the next λ_2 coordinates, and so on, then $X = S_n/S_{\lambda_1} \times \ldots \times S_{\lambda_k}$. The space $L(X)$ can be taken as all real valued functions on X because all irreducible representations are real. Thus $L(X) = M^\lambda$ of Chapter 7. The decomposition of $L(X)$ is given by Young's rule (see Chapter 7-B). Here are some special cases.

Case 1. $\lambda = (n-1, 1)$. This is simple choice data, people choosing one out of n items. The set X may be regarded as $\{1, 2, \ldots, n\}$ and $f(x)$ is the number of people choosing x. The decomposition is

$$L(X) = S^n \oplus S^{n-1,1}$$

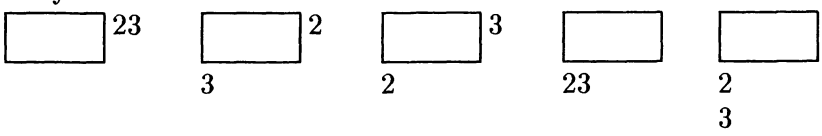
with S^n the trivial representation and $S^{n-1,1}$ the $n-1$ dimensional representation. This amounts to $f(i) = \overline{f} + (f(i) - \overline{f})$ with $\overline{f} = \frac{1}{n}\Sigma f(i)$.

Case 2. $\lambda = (n-2,1,1)$. Here people pick their favorite and second favorite items from a list of n (order matters). The space X may be regarded as $\{(i,j),\ 1 \le i \ne j \le n\}$ so $|X| = n(n-1)$. The decomposition of $L(X)$ is

$$L(X) = S^n \oplus 2S^{n-1,1} \oplus S^{n-2,2} \oplus S^{n-2,1,1}$$

$$\dim\ n(n-1)\ \ 1\ \ \ 2(n-1)\ \ n(n-3)/2\ \ (n-1)(n-2)/2.$$

To derive this, start with $\lambda = (n-2,1,1)$. Consider $n-2$ ones, one two, and one three. Young's rule asks for the number of ways of arranging these symbols into arrays which are increasing along rows, and strictly increasing down columns. The only ways are



where the block indicates $n-2$ ones. The number of times each shape appears gives the multiplicity of the corresponding irreducible. The dimensions are computed from the hook length formula of Chapter 7.

Here there is a natural interpretation of the decomposition. The projection into S^n represents the best constant approximation to f, i.e. $\overline{f} = \frac{1}{n(n-1)}\Sigma f(i,j)$. The two $S^{n-1,1}$ spaces give the effect of the first coordinate and the second coordinate respectively. The projection of f into $S^n \oplus 2S^{n-1,1}$ gives the best approximation to f of the form

$$a + f_1(i) + f_2(j) \text{ with } \Sigma f_1(i) = \Sigma f_2(j) = 0.$$

The projection into $S^{n-2,2}$ can be thought of as the best approximation of f, orthogonal to $S^n \oplus 2S^{n-1,1}$, of form

$$f_3\{i,j\}$$

the brackets indicating unordered pairs. The projection onto $S^{n-2,1,1}$ can be thought of as a residual, or what's left when the first four terms are subtracted off.

EXERCISE 3. Verify the decomposition

$$\begin{array}{lllllllll} M^{n-3,1,1,1} & = S^n & \oplus\ 3S^{n-1,1} & \oplus\ 3S^{n-2,2} & \oplus\ S^{n-3,3} & \oplus\ 2S^{n-3,2,1} & \oplus\ 3S^{n-2,1,1} & \oplus\ S^{n-3,1,1,1} \\ n(n-1)(n-2) & 1 & 3(n-1) & \frac{3n(n-3)}{2} & \frac{n(n-1)(n-5)}{6} & \frac{2n(n-2)(n-4)}{3} & 3\binom{n-1}{2} & \binom{n-1}{3} \end{array}$$

Can you give an interpretation to the subspaces involved?

3. *Computing projections.* We turn now to the problem of actually computing the projections, choosing bases, and so on.

Method 1 – *Character Theory.* Let G be a group, $\rho: G \to GL(V)$ a representation. The decomposition of V into irreducibles usually requires choosing a basis. There is a coarser direct sum decomposition that is basis free. Let $\chi_1, \chi_2, \ldots, \chi_h$ be the distinct characters of the irreducible representations $W_1, \ldots, W_h$ of G. Let

$d_1, d_2, \ldots, d_h$ be their degrees. Let $V = U_1 \oplus U_2 \oplus \ldots \oplus U_m$ be the decomposition of V into a direct sum of irreducible representations. For $i = 1, \ldots, h$, denote by V_i the direct sum of the spaces U_j which are isomorphic to W_i. Then

$$V = V_1 \oplus \ldots \oplus V_h.$$

This decomposition is canonical in the following sense:

Theorem 1.

(1) The decomposition $V = V_1 \oplus \ldots \oplus V_h$ does not depend on the initially chosen decomposition of V.

(2) The projection Π_i of V onto V_i associated with this decomposition is given by

$$\Pi_i = \frac{d_i}{|G|} \sum_{t \in G} \chi_i(t)^* \rho(t).$$

(3) If the original representation is unitary, then Π_i is an orthogonal projection.

Proof. We first prove (2). Restrict attention to an irreducible subspace U_j. Then ρ restricted to U_j is an irreducible representation with character χ say. Let Π_i be defined by the formula in (2). Since U_j is invariant, Π_i maps U_j into itself. Since Π_i and ρ commute, Π_i restricted to U_j is a constant times the identity. This constant is

$$< \chi_i | \chi >^* = \begin{cases} 0 & \text{if } \chi \neq \chi_i \\ 1 & \text{if } \chi \neq \chi_i. \end{cases}$$

This proves that Π_i is a projection onto V_i. Since Π_i does not depend on the originally chosen decomposition, (1) follows.

For (3), clearly Π_i is a projection. To show it is orthogonal, we must show that $\Pi_i^* = \Pi_i$. This is obvious if ρ is unitary. □

EXERCISE 4. Consider the voting data of Chapter 5-B. Compute the projections into the irreducible subspaces for the people voting for only two candidates (see Case 2 in Section 2 above. The characters of S_5 are in James and Kerber (1981)) compare with people voting for only one candidate.

We will find use for this theorem as a practical tool for computing projections. Here is a theoretical application which shows that the spectral analysis on groups developed in Section A is the same as the spectral analysis developed in this section.

COROLLARY 1. *Let R, $L(G)$ be the regular representation of the finite group G. Let ρ be an irreducible representation and V_ρ the direct sum of all irreducibles in $L(G)$ isomorphic to ρ. For $f \in L(G)$, the orthogonal projection of f into the space V_ρ is given by*

$$\Pi_\rho f(s) = \frac{d_\rho}{|G|} Tr(\rho(s^{-1}) \hat{f}(\rho)).$$

Proof. Using Theorem 1, we must show

$$\Pi_\rho f(s) = \frac{d_\rho}{|G|} \sum_\rho \chi_\rho(t^{-1}) R_t\ f(s) = \frac{d_\rho}{|G|} Tr\{\rho(s^{-1})\hat{f}(\rho)\}.$$

Both sides are linear in f, so take $f = \delta_u$. Cancelling common factors, the left side equals

$$\Sigma_t\ \chi_\rho(t^{-1})\delta_u(t^{-1}s) = \chi_\rho(us^{-1}) = Tr\{\rho(s^{-1})\rho(u)\} = Tr\{\rho(s^{-1})\hat{\delta}_u(\rho)\}.$$

The projection is orthogonal because the regular representation is unitary. Note that the right side of the formula doesn't depend on the basis chosen for ρ. □

Example. Take $G = S_n$, and X as the k sets of $\{1, 2, \ldots, n\}$ as discussed in example 2 of Chapter 5B. Here $X = S_n/S_k \times S_{n-k}$ and the representation decomposes without multiplicity (see Exercise 2 of Chapter 7) as

$$\begin{array}{lccccccc} L(X) & = & S^n & \oplus & S^{n-1,1} & \oplus & S^{n-2,2} & \oplus \ \ldots \ \oplus \ S^{n-k,k} \\ \dim \binom{n}{k} & & 1 & & n-1 & & \binom{n}{2} - \binom{n}{1} & \ldots \quad \binom{n}{k} - \binom{n}{k-1}. \end{array}$$

Here the projection given by Theorem 1 is simply the projection into irreducibles. This holds for data on any homogeneous space which is a Gelfand pair (see Chapter 3-F).

Example. Take $G = S_n$ and $X = S_n/S_{n-2} \times S_1 \times S_1$. Here there is multiplicity:

$$L(X) = S^n \oplus 2S^{n-1,1} \oplus S^{n-2,2} \oplus S^{n-2,1,1}.$$

A further decomposition of the projection into $2S^{n-1,1}$ is required. One way to do this is described in Exercises 2.8, 2.9 and 2.10 of Serre (1977, Sections 2.6 and 2.7). A second way to do it is outlined next.

Method 2 – Following known vectors. One problem with Theorem 1 is that it involves a sum over the group. For homogeneous spaces like $S_{49}/S_6 \times S_{43}$, this is simply not feasible.

Let G be a group and (ρ, V) a representation. Suppose V is equipped with an invariant inner product. We are often in a position where we know, or can guess at, vectors $w_1, w_2, \ldots, w_J \in V$ which generate an invariant subspace $W \subset V$. It is then straightforward to orthonormalize w_i using the Gram-Schmidt procedure, forming $w_1^*, w_2^*, \ldots, w_J^*$. Then the projection of v on W is

$$\Pi_W v = \Sigma < v, w_j^* > w_j^*.$$

Computations using this approach usually only involve a small fraction of work required for the character theory approach.

Example. Consider the decomposition arising from $X = S_n/S_{n-2} \times S_1 \times S_1$-ordered pairs out of n. The pieces are

$$L(X) = S^n \oplus 2S^{n-1,1} \oplus S^{n-2,2} \oplus S^{n-2,1,1}.$$

Here $L(X)$ is considered as the space of all real functions on ordered pairs with $< f, g > = \Sigma_{i,j} f(i,j)g(i,j)$. The $n-1$ functions $f_k(i,j) = \delta_k(i) - \frac{1}{n}$, $1 \leq k < n$ are linearly independent and span an $n-1$ dimensional subspace of $2S^{n-1,1}$. The $n-1$ functions $g_k(i,j) = \delta_k(j) - \frac{1}{n}, 1 \leq k < n$ are linearly independent of the f_k and each other. They span the rest of $2S^{n-1,1}$.

This gives a natural way of decomposing the remaining subspace in the example of the last section.

Remark 1. This approach is available for decomposing *any* of the subspaces M^λ. The relevant details are given in Section 17 of James (1978). In deriving a version of Young's rule that works for finite fields, James introduces a hierarchy of invariant subspaces $S^{\mu\#,\mu}$ which split M^λ into progressively more refined pieces (ending with irreducibles). He gives an explicit basis for each of these subspaces involving sums over the column stabilizer subgroups. These sums, and attendant computations, seem computationally manageable.

As an example, we know that each M^λ contains the irreducible S^λ once. Recall from Chapter 7, a standard Young tableau is a tableau which is decreasing across rows and down columns. Recall that for a tableau t the vector (or function or polytabloid) e_t is defined by

$$e_t = \sum_{\pi \in C_t} \operatorname{sgn}(\pi) e_{\pi\{t\}},$$

where $\{t\}$ is the tabloid associated to t, and C_t is the column stabilizer of t. James shows that $\{e_t : t$ is a standard Young tableau of shape $\lambda\}$ is a basis for S^λ in M^λ. For λ's without many parts, $|C_t|$ is manageable (e.g., it has size 2^k for $\lambda = (n-k, k)$). The number of standard Young tableaux is given by the hook length formula of Chapter 7. The e_t can be orthogonalized to e_t^*. Then the projection of $v \in M^\lambda$ can be computed through $v \cdot e_t^*$.

Remark 2. Often a subspace $W \subset V$ has a simple data analytic interpretation, but for computational convenience, the projections are computed with respect to a basis which scrambles things up. A second problem: the dimension of W can be smaller than the number of natural projections. For instance, the decomposition arising in studying unordered three sets out of six is

$$M^{3,3} = S^6 \oplus S^{5,1} \oplus S^{4,2} \oplus S^{3,3}$$

the dimension of $S^{4,2}$ is 9 $(= \binom{6}{2} - \binom{6}{1})$. Now $S^{4,2}$ is the space of unordered pair effects. There are $\binom{6}{2} = 15$ such effects it is natural to look at.

Colin Mallows has suggested a remedy for these problems. Given $v \in V$, project it to $v^* \in W$. Then take the natural vectors of interest, say $v_1, v_2, \ldots, v_\ell$, project them into $v_i^* \in W$, and plot $< v^*, v_i^* >$ versus i. This allows us to try to interpret the projections on a natural scale. It assumes v^*, v_i^* have been normalized to be unit vectors. See Diaconis (1989) for applications.

Method 3 – Radon transforms. There is a special technique available for decomposing the representations M^λ associated to partially ranked data of shape λ.

For clarity, this will be explained for the multiplicity free case $\lambda = (n-k, k)$ with $k \leq n/2$.

A vector $f \in M^{n-k,k}$ can be regarded as a function on k sets of an n set. For $1 \leq j \leq k$, define a mapping

$$R^+ : M^{n-j,j} \to M^{n-k,k} \text{ by } R^+ f(s) = \sum_{s \supset r} f(r)$$

where $|r| = j$, $|s| = k$. This Radon transform R^+ has an inverse $R^- : M^{n-k,k} \to M^{n-j,j}$ satisfying $R^- R^+ f = f$. An explicit form of the inverse is given by Graham, Li and Li (1980). If $M^{n-k,k}$ and $M^{n-j,j}$ are given bases consisting of delta functions on the k sets and j sets respectively, then the r, s element of R^- is

$$\frac{(-1)^{k-j}(k-j)}{(-1)^{|s-r|}|s-r|} \frac{1}{\binom{n-j}{|s-r|}}.$$

Note that both R^+ and R^- commute with the action of S_n. Composing maps in the other way, R^+R^- gives a map $M^{n-k,k} \to M^{n-k,k}$. This map is an orthogonal projection onto the single copy of $M^{n-j,j}$ contained in $M^{n-k,k}$:

LEMMA 1. *The map R^+R^- is an orthogonal projection on $M^{n-k,k}$ with range isomorphic to $M^{n-j,j}$.*

Proof. Since $R^+R^-R^+R^- = R^+(R^-R^+)R^- = R^+R^-$, the map is a projection. To show that it is orthogonal we must show $(R^+R^-)^t = R^+R^-$. Let s_1 and s_2 be k sets. The s_1, s_2 entry of R^+R^- is proportional to

$$\sum_{|r|=j} \delta_{rs_1} \frac{1}{(-1)^{|s_2-r|}|s_2-r|} \frac{1}{\binom{n-j}{|s_2-r|}}.$$

With $\delta_{rs} = 1$ or 0 as $r \subset s$ or not. This sum is a sum of terms of form

$$\frac{1}{(-1)^{k-\ell}(k-\ell)} \frac{1}{\binom{n-j}{k-\ell}},$$

the multiplicity of this term being $\#\{r \subset s_1 : r \cap s_2 = \ell\}$. This is a patently symmetric in s_1 and s_2, so orthogonality follows.

Finally, both R^+ and R^- commute with the action of S_n. The map R^- is onto $M^{n-j,j}$, so R^+ is an isomorphism of $M^{n-j,j}$ with the range of R^+R^- i.e. with range R^+. □

To use the lemma, define $R^+R^- = \pi_j$. Then $I - \pi_j$ is also an orthogonal projection. We know $M^{n-k,k} = S^n \oplus S^{n-1,1} \oplus S^{n-2-2} \oplus \ldots \oplus S^{n-k,k} = M^{n-j,j} \oplus S^{n-j-1,j+1} \oplus \ldots \oplus S^{n-k,k}$. One may thus proceed by taking $j = k-1, k-2, \ldots, 1$ inductively. This procedure is computationally feasiable for n large and k small.

A similar procedure is available for any partition λ using the results in Section 17 of James (1978). Diaconis (1987) gives details.

Final remarks on choice of basis. Return to data f on a group G. The Fourier transform $\hat{f}(\rho)$ can be a completely arbitrary matrix for general f. To see this, just define f by the inversion theorem to have a prescribed $\hat{f}(\rho)$. There *is* a rather complex restriction, akin to Bochner's theorem but more complex, when f is positive. For practical purposes $\hat{f}(\rho)$ is an essentially arbitrary matrix.

We have available the possibility of changing bases, changing from $\hat{f}(\rho)$ to $A\,\hat{f}(\rho)A^{-1}$. As A varies, the invariants of $\hat{f}(\rho)$ are its eigenvalues, or its "canonical form."

To compare coefficients within $\hat{f}(\rho)$ and between various ρ's, it seems natural to consider only unitary (or orthogonal) base changes. Then, there is not much that can be done to bring $\hat{f}(\rho)$ into a simple form. One possibility is to rotate to make the first few rows of $\hat{f}$ as large as possible.

Other possibilities are to change bases to bring the matrix into a simple form. For example, any matrix can be conjugated by an orthogonal matrix to the sum of a diagonal and skew symmetric matrix: $A = \frac{1}{2}(A + A^t) + \frac{1}{2}(A - A^t)$, and $A + A^t = \Gamma D \Gamma^t$, so $\Gamma A \Gamma^t = \frac{1}{2}D + S$, with D diagonal and S skew symmetric. Any matrix can be orthogonally conjugated to lower triangular form (essentially) see Golub and van Loane (1983).

The choice of bases must balance off

- computational convenience
- ease of interpretation
- convenience of orthogonality
- maximal informativeness of a few projections
- invariance considerations.

Nobody ever said statistics was easy.

At present, we can have the first three properties, for partially ranked data, by using methods 2 or 3 and Mallows' idea as outlined under method 2. There is plenty to think about.

C. Analysis of variance.

Analysis of variance (ANOVA) is a set of techniques for analyzing cross tabulated data such as two-way arrays. Data analytically, ANOVA is a special case of spectral analysis as described in Section B. This section develops the connection, and explains a technique introduced by Peter Fortini for defining the group for the array from a naturally given set of factors.

There is more to ANOVA than the data analytic aspects presented here. Section D-4 describes some of the other ideas. Hopefully they can be developed for other types of spectral analysis.

1. *Classical examples.*

Example 1. Two-way ANOVA with 1 *repetition per cell.* Here data are collected in a two way layout

x_{11}	x_{12}	...	x_{1J}
x_{21}			
$\vdots$			
x_{I1}			x_{IJ}

The rows represent the I different levels of one factor, the columns represent the J different levels of a second factor. The classical examples involve agriculture: the x_{ij} might represent the crop yield, the row factors might be different types of seed, the column factors might be different types of fertilizer. In a psychology example the data might be reaction time, the row factor might be wording of question, the column factor might be type of recording device.

There are two basic steps in analyzing such data: forming estimates of "main effects" and the testing ritual involving F statistics that is often called ANOVA. The first is simple and quite sensible: calculate the average of all numbers in the table $x_{..}$ – the grand mean. Subtract $x_{..}$ from x_{ij}. Calculate "row effects" $x_{i.} = \frac{1}{J}\Sigma_j\{x_{ij} - x_{..}\}$. Subtract the row effects and calculate column effects $x_{.j}$. Subtract the column effects forming residuals: $r_{ij} = x_{ij} - x_{..} - x_{i.} - x_{.j}$. This is attempting an additive decomposition of the data:

$$x_{ij} \doteq x_{..} + x_{i.} + x_{.j}.$$

If the residuals are small, the approximation makes a useful, simple description of the data. One then looks at the main effects – e.g. if the column effects are ordered, they might be plotted against j to see their shape, etc. The residuals can be plotted (does there seem to be a relation with $x_{i.}$ or $x_{.j}$?). If an additive fit fails, one can try different scales – an additive fit to log x_{ij}. Tukey (1977) vigorously illustrates this approach to ANOVA.

The second piece of analysis is a way of performing a statistical test to see if the main effects are zero. There are several different tests – all main effects zero, row effects zero, column effects zero, and even grand mean zero. These tests all have a geometric interpretation: they are ratios of the squared lengths of the "effect" vectors to the squared length of the residual vector. The tests can be justified in language involving normal distributions or in language involving permutations of the residuals.

I cannot hope to adequately review the statistical aspects of ANOVA here. The classical text by Scheffé (1959) is still most readable and useful. Two fine survey articles are those by Tjur (1984) and Speed (1987). Further references and discussion are given in Section D-4.

We now rephrase this as a spectral analysis problem. The data x_{ij} can be thought of as a function defined on the set of ordered pairs $X = \{(i,j)\ 1 \leq i \leq I,\ 1 \leq j \leq J\}$. The group $G = S_I \times S_J$ (where S_n is the symmetric group on n letters) acts transitively on the ordered pairs. This gives a representation of $S_I \times S_J$ on $L(X)$ which is just the product of the usual I dimensional representation we called $M^{I-1,1}$ of S_I with $M^{J-1,1}$ of S_1. Now $M^{I-1,1}$ splits into the constants, and an $I-1$ dimensional irreducible. We write $M^{I-1,1} = S^I \oplus S^{I-1,1}$. So

$$\begin{array}{lccccccc} & M^{I-1} \otimes M^{J-1} & = (S^I \oplus S^{I-1,1}) \otimes (S^J \oplus S^{J-1,1}) & = & Triv \oplus S^{I-1,1} & \otimes & Triv \oplus S^{J-1,1} & \otimes \; Triv \oplus S^{I-1,1} \otimes S^{J-1,1} \\ dim & I \times J & & & 1 & & I-1 & J-1 \qquad (I-1)(J-1) \end{array}$$

The observed array x_{ij} is just a vector in $L(X)$. The projection of x into the invariant irreducible subspaces gives the usual ANOVA decompositon: the projection onto the one dimensional space of constants is the array with constant entry $x_{..}$. The projection onto $S^{I-1,1} \otimes$ triv is onto all arrays which are constant in each row, with the constants summing to zero. These constants are $x_{i.}$. The final space is spanned by the residuals r_{ij}.

The lengths of these projections give the usual sums of squares required for the classical ANOVA table. The dimensions of the various irreducible subspaces give the degrees of freedom for the classical F tests.

Example 2. Two way ANOVA with repeated entries - wreath products. In this example, data are collected in a two way array, but k observations are collected at each level of the two factors. It is intuitively clear that the appropriate group of symmetries allows permutations of rows and columns and, within each row and column, an arbitrary permutation of the k elements in that cell among each other. To write this down neatly, it is useful to introduce the notion of wreath product.

Let G be any group and H a subgroup of S_n. Define a group G wr H as follows. The group consists of the set of elements in $G^n \times H = \{(g_1, g_2, \ldots, g_n;\ h)\}$. The product of two elements is

$$(g_1, \ldots, g_n;\ h)(g'_1, \ldots, g'_n;\ h') = (g_1 g'_{h^{-1}(1)}, \ldots, g_n g'_{h^{-1}(n)};\ hh').$$

The identity is $(i_G, \ldots, i_G;\ i_H)$, $(g_1, \ldots, g_n;\ h)^{-1} = (g^{-1}_{h^{-1}(1)}, \ldots, g^{-1}_{h^{-1}(n)};\ h^{-1})$. The subgroup G^n sits in G wr H as a normal subgroup. H sits in G wr H as $\{(i_G, \ldots, i_G;\ h)\}$. G wr H is the semi-direct product of G^n and H.

For an $I \times J$ table with k replications per cell, the natural symmetry group is S_k wr$(S_I \times S_J)$. The representation theory of wreath products is completely known in terms of the irreducible representations of G and H. A reasonable treatment is Chapter 4 of James and Kerber (1981). When this theory is applied to ANOVA there are no surprises; the decomposition is the one derived in all standard texts; see Example 4 below for a special case.

Example 3. The Diallel Cross Design. This design is used by animal or plant breeders to find new hybrid species of known species. Consider n known types of plant. Form all $\binom{n}{2}$ crosses of distinct strains. The data are some measurable characteristic of each offspring (yield, size, etc.). Here the data are indexed by unordered pairs $\{i, j\}$. The natural symmetry group is S_n; the standard ANOVA decomposition corresponds to the decomposition of what we have called $M^{n-2,2} \cong S^n \oplus S^{n-1,1} \oplus S^{n-2,2}$.

We will return to examples after a general definition of the symmetry group of a designed experiment.

2. *A systematic approach.* In the examples above we just wrote down a group and showed how spectral analysis gave "the usual" ANOVA.

Alan James (1957, 1982) and Ted Hannan (1965) introduced ideas to relate an algebraic object to a designed experiment. Gleason and McGilchrist (1978) give

a more detailed account of these ideas. Lederman (1967) independently pointed out the connection between ANOVA and representation theory.

Peter Fortini (1977) extended these ideas to give a reasonably systematic theory. His work associates a group in a natural way to an experiment given in terms of classical factors. Then spectral analysis can be used as before. Here is a brief account of Fortini's ideas.

Let X be a finite set. We will work with observations indexed by X.

Definition 1. A *factor* f of the design indexed by X is a set valued map from X to the elements of a finite set F. If $|X| = N$, $|F| = k$, a factor can be described by an $N \times k$ matrix f where $f_{x\ell} = 1$ if $\ell \in f(x)$, $f_{x\ell} = 0$ otherwise. A *multifactorial design* is a set X and a collection of factors $(f_1, F_1), \ldots, (f_k, F_k)$.

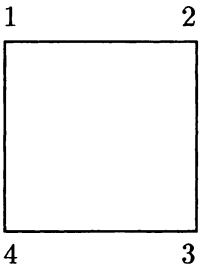
Definition 2. The *automorphism group* G of a multifactorial design is the group of all permutations g of X with the property that for each factor f_i there is a permutation g_i of F_i such that $f_i(gx) = g_i(f_i(x))$.

Thus, if x is associated with levels $f_i(x)$, gx is associated to levels $g_i f_i(x)$. The *outcome set* V of a numerical experiment indexed by X is the set of all real valued functions on X. This has basis e_x, and we have the permutation representation of G acting on V. An *analysis of variance* is a decomposition of V into a direct sum of irreducible invariant subspaces V_i. The projections of $v \in V$ onto V_i are called main effects. The squared lengths of these projections form the ith line of an ANOVA table: the number of degrees of freedom of the ith line is dim V_i.

Example 4. Two treatments with two objects per treatment. To understand the definitions, consider comparing two treatments: A or B with two objects for each treatment. Take X to be the set of four objects, labeled $\{1, 2, 3, 4\}$. There is one factor – did the object get treatment A or B? Suppose that objects 1 and 3 get treatment A and objects 2 and 4 get treatment B. The factor matrix is

	A	B
1	1	0
2	0	1
3	1	0
4	0	1

The permutations of X that are in G are id, (1 3), (2 4), (1 3)(2 4), each associated with the identity permutation of factors A and B, and (1 2)(3 4), (1 4)(2 3), (1 2 3 4), (4 3 2 1) each associated with transposing A and B. We thus get an eight element automorphism group. Observe that the group is the symmetry group of the square



It is often called D_4 – the dihedral group on four letters. Observe too that if we think about the problem directly, as we did before, the "obvious" symmetry group is S_2 wr S_2. This is an eight element group which is isomorphic to D_4.

How does the four-dimensional space V decompose? Here, the decomposition is obvious, but as an exercise, we follow Fortini and derive the result using character theory. The character table for D_4 is in Chapter 5 of Serre (1977):

Class	χ_1	χ_2	χ_3	χ_4	χ_5	Perm rep.
1	1	1	2	1	1	4
(1 3)(2 4)	1	1	-2	1	1	0
(1 2)(3 4), (1 4)(2 3)	1	-1	0	-1	1	0
(1 3), (2 4)	1	-1	0	1	-1	2
(1 2 3 4), (1 4 2 3)	1	1	0	-1	-1	0

Looking across the first row, there are 5 distinct irreducible representations, four of dimension 1 and one of dimension 2. Which of these appear in the permutation representation? The character of the permutation representation is the number of fixed points of G acting on X. This is given in the last column above. The multiplicity of χ_i in the permutation representation is given by

$$m_i = < \chi_i | \chi_{\text{perm}} > = \frac{1}{8} \Sigma \, \chi_i(g) \chi_{\text{perm}}(g).$$

We get $m_1 = 1$, $m_2 = 0$, $m_3 = 1$, $m_4 = 1, m_5 = 0$. Thus, $V = V_1 \oplus V_2 \oplus V_3$. V_1 is the 1 dimensional grand mean space, V_2 is the space: $\{y \in V : (y_1 + y_3) - (y_2 + y_4) = 0\}$. It represents the difference between the average of the group means. The space V_3 is a 2 dimensional space of "what's left over."

Example 3. revisited. The need for factors taking more than one value on a given observation is well illustrated by the diallel cross. Here $X = \{\{i, j\};\ i \neq j\}$. $F = \{1, 2, \ldots, n\}$, and $f\{i, j\} = \{i, j\}$. Thus the matrix has two ones in each row. The automorphism group is a subgroup of $S_{\binom{n}{2}}$. A bit of reflection shows that it is S_n, and that V is what we have been calling $M^{n-2,2}$.

Example 5. Balanced incomplete blocks. Let us begin with an example taken from Cochran and Cox (1957, pg. 443) on the effects of aging on the tenderness of beef. Six periods of storage (0, 1, 2, 4, 9, and 18 days) were tested. These treatments are denoted 1, 2, 3, 4, 5, 6 respectively.

To facilitate comparison, roasts from symmetric locations (left/right sides) were paired into blocks of size 2. There are $15 = \binom{6}{2}$ ways to treat a pair. Scoring was done by 4 judges, each marking on a scale from 0 to 10. The data shows their totals out of 40, a high score indicating very tender beef.

Block	1	2	3	4	5	6	7	8
Pair	{1,2}	{3,4}	{5,6}	{1,3}	{2,5}	{4,6}	{1,4}	{2,6}
Scores	7 17	26 25	33 29	17 27	23 27	29 30	10 25	26 37
Block Sums	24	51	62	44	50	59	35	63

Block	9	10	11	12	13	14	15
Pair	{3,5}	{1,5}	{2,4}	{3,6}	{1,6}	{2,3}	{4,5}
Scores	24 26	25 40	25 34	34 32	11 27	24 21	26 32
Block Sums	50	65	59	66	38	45	58

Thus block 1 was given treatments 1 and 2. The roast receiving treatment 1 was rated at 7. The roast receiving treatment 2 was rated 17.

The first thoughts for such an analysis run as follows: It is natural to compute the sum for all roasts receiving treatment i:

Treatment	1	2	3	4	5	6
Sum	70	115	132	139	158	155

This suggests longer aging improves tenderness, the effect perhaps peaking at treatment 5 (9 days).

Now it is natural to try to see if there is a block effect: roasts from different location vary in tenderness, and if one of the treatments was tried on more tender blocks, this would favor the treatment.

The natural adjustment subtracts from the ith treatment total the sum of the block averages (here block sum/2) for the blocks containing the ith treatment. Chapter 5 of Scheffé (1959) gives a careful, clear description. In the example, the adjusted sums are

Treatment	1	2	3	4	5	6
Adjusted Sum	-33	-5.5	4	8	15.5	11

These now sum to zero.

Here, the adjustment doesn't affect the conclusions drawn above. We leave further details to Cochran and Cox who carry out the usual analysis and conclude that storage up to about a week increases tenderness.

In general, there are t treatments to be compared in blocks of size $k < t$. An *unreduced balanced incomplete block design* involves all possible $\binom{t}{k}$ blocks and so $k\binom{t}{k}$ basic units (in the example, $t = 6$, $k = 2$ and 30 roasts were involved).

One natural way to index such data is as a pair (i,s) with $1 \leq i \leq t$ denoting the treatment applied to that unit, and s of cardinality $|s| = k - 1$ denoting the subset in the same block as the given unit. Thus $X = \{(i, s)\}$ and $|X| = t\binom{t-1}{k-1} = k\binom{t}{k}$.

The group S_t that permutes treatment labels operates transitively on X, and we see that $L(X)$ can be identified with $M^{t-k,k-1,1}$ of Chapter 7. Another way to arrive at this result begins with the idea that this experiment has two factors: "treatments" $f_1(i, s) = i$, and "blocks" $f_2\{i, s\} = \{i \cup s\}$.

What automorphisms are possible? A little reflection shows that only permutations that permute treatments among themselves are allowed; treatments that permute things within a block are ruled out and allowable treatments that move blocks around can be achieved by permuting treatments. Thus the automorphism group is isomorphic to S_t.

The decomposition now follows from Young's rule. Before stating the result, let us examine the introductory example on aging of meat. Here $t = 6, k = 2$, and we have

$$M^{4,1,1} = S^6 \oplus 2S^{5,1} \oplus S^{4,2} \oplus S^{4,1,1}$$
$$\dim\ 30 \quad 1 \quad 2 \times 5 \quad 9 \quad 10$$

(see Case 2 of Section B-2 above). The projection onto the one dimensional space S^6 is the grand mean. The treatment effects can be identified as one of the $S^{5,1}$ spaces. The block effects space, orthogonal to $S^6 \oplus S^{5,1}$, is $S^{4,2}$ (the full 15-dimensional block effect space $M^{4,2} = S^6 \oplus S^{5,1} \oplus S^{4,2}$). These projections constitute the classical analysis for this experiment, projecting into treatment and blocks adjusted for treatment.

The remaining $S^{5,1}$ gives a new piece of analysis due to Fortini. To give it a clear interpretation let us change the imagery. Suppose there are 6 amounts of vanilla added to ice cream (from none to a fair amount). Fifteen people each taste two servings each in a balanced incomplete block involving 30 servings in all. The treatment effects are interpreted as before. The block effects become subject effects – some people are systematically higher than others.

The classical analysis assumes that treatment and block effects are additive. However, this would fail if tasters give ratings partially by comparison. The 2nd copy of $S^{5,1}$ can be interpreted as the additive effect on the rating of treatment i due to being in the same block with treatment j.

In tasting examples, this (and its higher order analogs described below) makes perfect sense. In the aging/tenderness experiment it seems perfectly sensible to set this effect to zero as is done classically.

To decompose $M^{t-k,k-1,1}$ in general, Young's rule begins with $t - k$ ones, $k - 1$ twos, and one 3. These are placed into any arrangement as a tableaux which is non-decreasing in rows and strictly increasing in columns. Thus

$$M^{t-k,k-1,1} = S^t \bigoplus_{j=1}^{k-1} 2S^{t-j,j} \oplus S^{t-k,k} \bigoplus_{j=1}^{k-1} S^{t-j-1,j,1}.$$

It being understood that any improper partition above doesn't contribute.

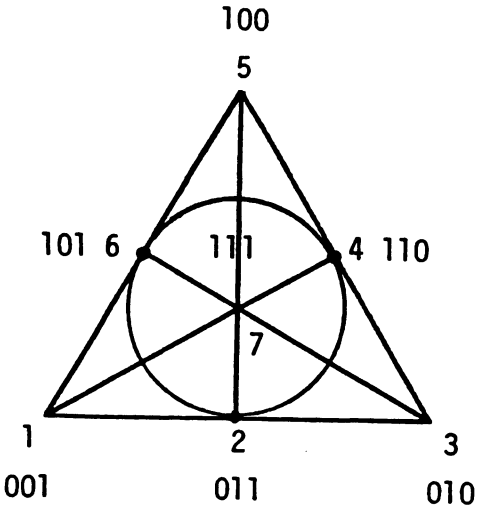
Fortini (1977) used this decomposition to build a class of standard linear models for which the decomposition gives the appropriate ANOVA. Calvin (1954) seems to be the first paper to extend the classical model to try to cope with this kind of non-additivity.

Example 6. The projective plane. As a final topic for this section we recall that there are many further types of block designs. For example, to compare 7 treatments, with block size 3, the following classical design can be used:

$$\{123\}\ \{345\}\ \{156\}\ \{246\}\ \{147\}\ \{257\}\ \{367\}.$$

Here, each pair appears together in exactly one block, so the same type of block/treatment analysis is available.

This design is constructed as the set of lines through the origan in the vector space Z_2^3. Each plane contains 3 points which gives rise to the blocks. These are edges in the figure below.



The group of this design is $GL_3(Z_2)$, a well studied simple group of order 168. For someone wishing to learn a little group theory, decomposing $L(X)$ makes a nice exercise. $L(X)$ is 21-dimensional. It decomposes into 4 irreducible pieces: the trivial representation, a 6-dimensional space of treatment effects, a 6-dimensional space of block effects, and an 8 dimensional irreducible space of residuals.

EXERCISE 5. Prove the assertions about $L(X)$. Hints: argue that $GL_3(Z_2)$ acts doubly transitively on lines and planes in Z_2^3. Using the "useful fact" in Chapter 7A, this gives two irreducible 6-dimensional representations of GL_3. Now consider $X = \{(i,p)\colon i \text{ a point, } p \text{ a plane, } i \in p\}$. Show that GL_3 acting on $X \times X$ has 4 orbits. Now use the "useful fact" of Chapter 7 to argue that $L(X)$ decomposes into 4 irreducibles.

This example is the tip of the iceberg called combinatorial design. Lander (1982) is a nice reference for the construction of block designs. Beth, Jungnichel, and Lenz (1986) is a recent encyclopedic reference. It seems like a worthwhile project to go through the classical designs, compute automorphisim groups, and compare the spectral decomposition with classical ANOVA.

D. Thoughts about spectral analysis.

This chapter builds on two well established traditions: the analysis of variance and the spectral analysis of time series. The new applications and variations of classical areas such as block designs should be regarded speculatively.

There is much of value in the parent theories as they appear in modern practice which I have not thought through in sufficient detail to carry over. This section outlines some of these techniques.

1. *Why projections onto invariant subspaces?* It is an empirical fact that the decomposition of data indexed by X into irreducible pieces is sometimes scientifically useful. Let me make an attempt to explain this. There is no one right way to analyze data. The spectral decomposition presents certain averages of the data which often seem sensible. The set of all averages is a 1-1 function of the data, so nothing is lost. Often the specific labeling of X is fairly arbitrary, and we would rather not have our conclusions depend on this labeling. In other circumstances, the order is worth taking into account, but it is useful to separate the part of the analysis that depends on the order from the part that is invariant under G.

Data indexed by X are represented by a function on X. Sometimes this function takes values in the integers, as in the case of counted data: "how many people in the population chose X?" Even here we want to be able to talk about averages and differences of averages, so we need to consider functions from X into the rationals Q. For the permutation group, the splitting of a representation over Q is the same as its splitting over the real or complex numbers.

Our function thus naturally sits in $L(X)$: the set of all complex valued functions. A subset of $L(X)$ can be thought of as a partial description. In practice these will be things like the constants or "first order functions;" sets of functions that are simple to describe or interpret. If the actual f can be well approximated by a simple f we regard this as useful. If f is "first order" and g is "first order" it seems natural that $\frac{1}{m}f$ or $f+g$ be the "first order" since, for example, $f+g$ has the interpretation of combining the two data sets (for counted data). This suggests that a basic partial description be a subspace of $L(X)$.

Finally, if G acts on X we want to consider descriptions that don't depend on the labelling: it seems unnatural that $f(x)$ could be "first order" but not $f(gx)$. Thus a natural descriptive unit is an invariant subspace of $L(X)$. Then *spectral analysis* is the decomposition of $f(x)$ into its projections on the irreducible invariant subspaces of $L(X)$, and the approximation of $f(x)$ by as small a number of its projections as give a reasonable fit.

Of course, this kind of linear analysis is just a start. Non-linear analyses may also be most useful (see Section (4) below). These are easiest to interpret if they have a simple connection to the linear theory (log f may be well approximated

by a simple linear fit).

2. *On the choice of group.* Spectral analysis begins with a group G acting on a set X. In some areas, like ANOVA, there is some theory to guide the choice of G. In other areas, like time series, physics dictates the choice. In new applications, there may be several possible groups among which to choose.

Clearly if G acts on X and H is a subgroup of G, the splitting of $L(X)$ into H irreducibles is finer than the splitting under G. Choosing the smallest group that preserves essential structure gives the largest number of projections. Preliminary analysis may start with larger groups.

Consider a two way $I \times J$ array. The group that gives classical ANOVA is $S_I \times S_J$. Another group that operates transitively is $Z_I \times Z_J$, operating by cyclically shifting each coordinate. A third transitive action is given by $S_I \times Z_J$. Any of these groups might be appropriate. Z_J preserves adjacency (or time order) while S_I invariance says adjacency isn't relevant. Thus $S_I \times Z_J$ might be deemed natural for a problem in which the rows index types of bird, the columns index months of the year, and the entries are number of birds sighted in that month by a local bird watching society.

EXERCISE 6. Derive the appropriate spectral analysis for $S_I \times Z_J$.

As a second example, consider $X = Z_2^k$. Four groups act naturally on this space: Z_2^k, $GL_k(Z_2)$, Z_2 wr S_k, and Z_2 wr Z_k. We consider these in turn.

a) Under Z_2^k, $L(X) = \bigoplus_{y \in X} V_y$ where V_y is the one dimensional space spanned by $x \to (-1)^{x \cdot y}$. Here spectral analysis amounts to the Fourier inversion theorem

$$f(x) = \frac{1}{2^k} \sum_y (-1)^{x \cdot y} \hat{f}(y).$$

b) $GL_k(Z_2)$ acts doubly transitively on X, so by Serre's exercise (2.6),

$$L(X) = V_0 \oplus V_1$$

with V_0 the constants and $V_1 = \{f : \Sigma f(x) = 0\}$.

c) Z_2 wr S_k (see Section C for notation) is the group of pairs (y, π) with $\pi \in S_k$, $y \in Z_2^k$ acting on X by $(y, \pi)x = (x_{\pi^{-1}(1)} + y_1, \ldots, x_{\pi^{-1}(k)} + y_k)$: you permute the coordinates by π and add y. It is straightforward to show that

$$L(X) = V_0 \oplus V_1 \oplus \ldots \oplus V_k$$

with V_j the linear span of $(-1)^{x \cdot y}$ for $|y| = j$. Thus spectral analysis under Z_2 wr S_k lumps together pieces of the spectral analysis under $G = Z_2^k$.

EXERCISE 7. Show that $L(X)$ splits as shown for Z_2 wr S_k. Find the decomposition with respect to Z_2 wr Z_k.

3. *On probability.* The data analytic approach to spectral analysis presented in this chapter is not based on probability. Spectral analysis is a reasonable, useful activity if f is not a sample, rather a complete enumeration of a population. Thus,

nothing is unknown, but there may still be plenty to do in condensing the data so that simple descriptions can be given. This point of view follows Tukey (1977).

Of course, there are many opportunities for probability to play a role. When we informally assess goodness of fit in a linear approximation such as $f(i,j) \doteq a + b_i + c_j$, we compare the residuals $r_{ij} = f(i,j) - a - b_i - c_j$, with the main effects a, b_i, c_j. If residuals are small, we regard the approximation as useful. The standard tests and confidence interval procedures are formalizations of this idea.

There is much more to do in constructing a believable probabilistic theory for spectral analysis. For example, the data $f(x)$ might be a sample from a larger population $F(x)$. If the sample size is small, there is no reasonable hope of estimating F, but one can hope to estimate some simple functionals of F such as its projections onto a few subspaces of interest. How the dimensionality of the subspaces should relate to the size of X and the sample size seems like a rich interesting question. Finch (1969) makes a start on these questions. See also the discussion in Section 3 of Diaconis (1985).

In some instances the bootstrap offers a reasonable way to wiggle data a bit. There are two obvious ways to bootstrap – sample iid from $f(x)$ to get $f^1(x)$, $f^2(x), \ldots, f^b(x)$, or fit a model and bootstrap residuals. Freedman (1981) describes these alternatives in regression problems. Of course, all of this is tied to some sampling like story. Without a believable sampling justification (as in complete enumeration problems) I find the bootstrap far too violent a perturbation of the data. Diaconis (1983) suggests less drastic perturbations.

I have been impressed with the usefulness of the basic normal perturbation model outlined in Section A-3. This serves as an accurate approximation for all sorts of schemes for quantifying "if the data had come out different, how would my conclusions differ?" or "if I had more data, how close is my best guess likely to be?"

Some further discussion of the need for probability is given later. It is worth emphasizing that spectral analysis can be useful without underlying probabilistic structure.

4. *Lessons from ANOVA.* A basic component of ANOVA is a test to see if projection of a given function onto a given subspace can be assumed to be zero. The standard test involves comparison of the observed projection with an estimate of "ambient noise" – usually the normalized length of the projection of the data onto the space of residuals. If the ratio of lengths is "large", then the projection cannot be asserted to be zero. Usually, "large" is quantified under the normal perturbation model of Section A-3.

Terry Speed (1987) gives an elegant survey of this material delineating a natural class of perturbation models involving patterned covariance matrices, for which the orthogonal projections in fact give the maximum likelihood estimates.

An important idea from classical ANOVA deals with fixed versus random effects. As data analytic motivation, consider a two way array with one observation per cell. Suppose the rows are different brands of typewriters and the columns are different typists. The x_{ij} might be average speed. If we are trying to evaluate typewriters, then the row averages are basic. It may be that the typists are thought of as drawn from a pool of typists. Then, the column averages are not of

particular interest – they are thought of as random effects. Their mean, variance, or distribution may be of interest as information about the population of typists. Tukey (1961) contains an interesting discussion of the distinction between fixed and random effects and its impact on ANOVA and time series data.

Modern ANOVA employs a wealth of non-linear techniques. These begin with transformations of x_{ij} to $T(x_{ij})$ aiming for linearity. Box and Cox (1964) or Hoaglin, Mosteller, and Tukey (1983, Chapter 8) give the basics. This is often supplemented by fitting more complex models such as

$$f(i,j) \doteq a + b_i + c_j + db_ic_j$$

as in Tukey (1949) or Mandel (1961). Stein (1966) gives a decision theoretic version.

More aggressive transformation techniques involving splines or more complex smoothers appear in approaches like projection pursuit (see, e.g. Huber (1985) or Henry (1983)) or Ace (see Breiman and Friedman (1985), Stone (1985), or Hastie and Tibshirani (1986)). None of these techniques are well tried out in ANOVA settings, but all seem worth thinking about and extending to other instances of spectral analysis.

Another important aspect of modern statistics is a concern for robustness – it may be that observed data is well approximated by a linear fit except for a few "wild values." For methods based on linear projections, even one wild value can foul up everything, making it seem as if no linear fit works.

One approach to these problems is to replace the usual averaging operators by robust versions such as medians. Hoaglin, Mosteller and Tukey (1985, Chapters 2-5) contains a good review. It is not at all clear if these approaches can be adapted to more complex designs which really lean on the additivity. Another approach is to try to remove or down-weight the outliers. Other approaches involve using perturbation distributions that are longer tailed than normal. Pasta (1987) contains a review of the problems and available remedies. It is fair to say that even for ANOVA this is a difficult problem, on the cutting edge of current research.

Here is another contribution of modern statistical theory to ANOVA. We now realize that while the projections have many appealing properties, they are not the most accurate estimates of population projections under the usual normal model. Non-linear shrinkage estimators can do better than the classical linear estimators. Stein (1966) shows how to modify the usual estimators in orthogonal designs to get improvement. I do not know of a systematic account for more general designs.

The Bayesian analysis of ANOVA data is important on its own – we often know quite a bit and want a way to combine this prior knowledge with the observations in a sensible way. Box and Tiao (1973) cover the basics. Consonni and Dawid (1985) give group related theory.

The Bayesian techniques give a reasonable way to get shrinkage estimates, possibly using conjugate priors and empirical Bayes ideas as surveyed by Morris (1983). Berger (1985, Chapter 4) surveys the recent literature on estimating a

normal mean. Again, most of these ideas are applicable in ANOVA and other spectral analysis problems but the details need to be worked out on a case by case basis.

An important problem that must be dealt with is missing data. The algebraic theory of ANOVA works neatly if things are neatly balanced. With a two-way layout and unequal numbers of observations per cell, there are no longer simple formulas for many quantities of interest. It is not unusual to have symmetry broken by having some subjects get sick or be lost for other reasons.

A rich collection of techniques has evolved for dealing with these problems. A most useful tool for computing "projections" has evolved as the EM algorithm of Dempster, Laird and Rubin (1977). They survey the literature. Dodge (1985) gives some special techniques for ANOVA problems.

There is some very recent work which relates the approach taken in this chapter to some of the mathematics of Chapter 3-F. In working with stochastic models for designed experiments, many workers have emphasized the importance and utility of "general balance." This is a condition introduced by Nelder (1965) which links the "design" part of the experiment to the structure of the assumed model. When the condition holds, all sorts of elegant formulas for computing best linear unbiased estimates are available. A recent survey of this material is given by Houtman and Speed (1983).

Recently, Bailey and Rowley (1986) have given a useful sufficient condition on the group of symmetries of the design (the definitions are a bit different than those in Section C) which forces generalized balance. There are too many details needed to state their results carefully. Very roughly, the group of symmetries acts on a space T of treatments, and the representation $L(T)$ must be multiplicity free (see Chapter 3F). The Bailey and Rowley paper is very clearly written and connects wonderfully with the material in this monograph.

5. *Lessons from time series.* Spectral analysis of time series and other signals is widely applied and has an enormous literature. Robinson (1982) surveys its history. It differs from ANOVA in having numerous hard science manifestations. Indeed, in some speech, geophysics, and signal processing applications the spectrum is naturally observed. Brillinger (1988) contains a list of such applications.

It is natural to try to emulate the success stories of spectral analysis for other groups. It seems fair to admit that at present there is no underlying "physical justification" for the spectral decomposition in other cases. None the less, it seems promising to try to plug into the experience and imagery of the established theory.

For example, it is well known that the periodogram (essentially the squared discrete Fourier transform) is not an accurate estimate of the amplitude of the Fourier transform at given frequency. It is asymptotically an exponential variable, and so not consistent as time goes on. Under smoothness assumptions on the underlying spectrum, smoothing the periodogram becomes sensible and leads to reasonable estimators. Olshen (1967) gives details.

The same problem is present in other spectral analyses. Under the normal perturbation model the Fourier transform has independent coordinates with fixed variance (see Example A-3). This continues to hold under broad alternative perturbation models. Without repeated observations, there is a limit to the accuracy

of estimates. Thought of in the ANOVA context this is not surprising – we can only hope for an accurate indicator of variability. We can think of formulating natural smoothness assumptions to allow increased accuracy.

When applicable, the sampling model is a natural replacement for signal plus noise – the Fourier transform applied to a population is simply a 1-1 transform and so gives a natural parameter to estimate.

Time series has goals in addition to estimating the spectrum at a point. In continuous time problems, there can be a continuous spectrum, and many other functionals are relevant. Tukey (1986) emphasizes this aspect and points to properties of ensembles of time series that are of scientific interest. For ranked data, there is an obvious prediction problem: how will a complete enumeration (or larger sample), turn out? More in line with time series, prediction might be this: people partially rank several quantities; later their true preferences are revealed. How good a guess can be made based on current knowledge?

Time series analysis has a time domain side which is based on models for the observed process. There is a healthy interplay between time and frequency domains. This is presented by Anderson (1971) or Priestly (1981). As with ANOVA, it's nice to see natural projections have an interpretation as reasonable estimates of parameters in a model. Models for data on homogeneous spaces are introduced in the next chapter. There hasn't been any serious work on the interplay with spectral analysis.

There are many modern versions of spectral analysis of time series. There are remedies for all of the problems discussed in the ANOVA section. It seems impossible to survey this briefly. Anderson (1971), Brillinger (1975) and Priestly (1981) are good references; leading to others. With work, most of these ideas should be transferrable to other spectral domains when the need arises.

One final thought: real examples in hard science settings are a sparkling part of time series analysis. If progress is to be made in generalization, similar examples will have to be found.