Graphical Display of Analysis of Variance with the Boxplot Matrix

RICHARD M. HEIBERGER

When the analysis of cross-classified data with multiple observations in each cell indicates significant interactions, we often need a graphical display to help in their interpretation. In this paper we construct a matrix of boxplots, one boxplot per cell, to combine in one visual image the crossing structure of the cell definitions with the within-cell detail on level and variability afforded by the boxplot. A series of boxplot matrices, with one boxplot matrix per line of the ANOVA table, can make quite complex patterns in the interactions of several factors become visually apparent. We illustrate and interpret several classical datasets, including some based on multi-factor designs. We give examples with blocking, covariates, high-order interaction, non-homogeneity of variance, and multiple error terms. We review several other graphical techniques that have been used in the analysis of cross-classified data. We discuss some of the technical details of building boxplot matrices and make available an implementation in each of two different graphical settings, *S* and XLISP-STAT.

KEY WORDS: Graphics; Boxplot; ANOVA; Interaction; *S*; XLISP-STAT.

1. INTRODUCTION

Sets of boxplot matrices can provide the needed practical graphic support for investigation of factorial and other complex designs. They combine the visual impact of a well-designed graphical display with the information customarily presented in the ANOVA table and in supplementary tables of cell means and cell standard deviations. They can provide a graphic presentation format that allows both the analyst and the client to

1. represent the raw data
2. represent the fitted model
3. display the individual lines of the ANOVA table
4. decide if the data look normal
5. decide if the variances are homogeneous
6. detect systematic patterns in non-homogeneity of variance
7. determine whether there are significant differences in the location attributable to the levels of the different factors
8. determine visually whether there are significant interactions between the different factors
9. illustrate higher-order interactions
10. display multiple error terms
11. display partial confounding

All but the last two items are illustrated in the examples in this paper.

Boxplot matrices for cross-classified data are easy to read because they track the structure of the data. The pattern of observed and missing cells is immediately visible. Each individual boxplot within the matrix can show asymmetry, spread, and outliers. Visual comparisons of the appearance of the set of boxplots in the matrix are immediate. Visual recognition of the correlation between the levels of the factors and the appearance of the boxplots is easy.

Boxplot matrices have high information content even though they are constructed with relatively low-tech graphics. Boxplot matrices are easily explained to a client.

Section 2 introduces the boxplot matrix by using it to display, and help in the analysis of, systematic non-homogeneity of variance in a two-way setting. We compare it to other commonly used techniques for displaying non-homogeneity. A series of boxplot matrices is used to illustrate each line of the analysis of variance table of suitably transformed data.

Section 3 discusses the mathematics behind the series of boxplot matrices. The boxplot matrices are shown to display projections in the n-dimensional dual space underlying least squares analysis. The display is in a visualization format suitable for ordinary clients (and perhaps statisticians).

Section 4 shows the boxplot matrices for each line of an analysis of covariance of data collected from a design more complex than the two-way layout of Section 2.

Section 5 discusses extensions to higher-order designs, customization of a series of plots, and modifications to the boxplot matrix that make it possible to display data matrices.

Section 6 gives a short history of the development of the boxplot matrix. It includes comments on the effect of implementing it two different graphic environments, *S* and XLISP-STAT. Access to the implementations available in statlib is described in Section 7. The appendix describes some of the implementation detail that was addressed while programming the function.

2. SYSTEMATIC NON-HOMOGENEITY OF VARIANCE

Figure 1 shows a boxplot matrix of the raw data on smoothness of paper [Mandel (1964, p. 325), repeated in Emerson (1991)]. The response variable, smoothness of paper, is an important measure in the manufacture of papers for the printing industry. The data were originally presented as part of the discussion of the Bekk method of measuring smoothness. The data are from a multi-laboratory reproducibility experiment. Five different types of paper (Material) were measured by four different laboratories (Lab). Each lab made 8 measurements of each material.

The rows and columns of the plot show the factorial structure of the data. A separate boxplot of the response variable appears within each Lab-Material cell. From this single plot of the data we are able to see much of the initial analysis of the data. The single most evident feature of this plot is that the within-cell variability increases with the response level. There is also a clear Material effect: the level and variability both increase with the Material index in all laboratories. There are some smaller differences between Labs and the interaction of Lab and Material: the response levels of Material 4 are lower than Material 5 in Labs 2, 3, and 4, and about equal in Lab 1.

The next step in the analysis is a data transformation designed to improve homogeneity of variance. The spread-versus-level plot in Figure 2, in

Table 1. ANOVA table for smoothness of paper in $\log_{10}$ units.

Source	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
lab	3	0.1961	0.0654	27.37	0.0000000
material	4	54.8091	13.7023	5736.97	0.0000000
lab:material	12	0.1393	0.0116	4.86	0.0000012
Residuals	140	0.3344	0.0024		

this case of $\log_{10}(\textsf{cell standard deviation})$ against $\log_{10}(\textsf{cell mean})$, has a slope of 1.1318, suggesting a transformation to the power $1 - 1.1318 = -0.1318$. Since -0.1318 is close to 0, we use the log transformation.

The logarithm of the original data is plotted in a boxplot matrix in Figure 3(a). There is no longer an obvious systematic variability in the spreads. Ordinary analysis of variance is the next reasonable step. The ANOVA table of the logarithms is displayed in Table 1. Each of its lines is illustrated in its own boxplot matrix in Figures 3(d), 3(e), 3(f) and 3(c). These four boxplot matrices provide a graphic partitioning of Figure 3(a). The main effects and interaction are all visible in these plots, each of which has been given its own scaling.

Comparison with Earlier Graphical Presentations

The graphical analysis of the Bekk data by the series of boxplot matrices needs to be compared to the earlier graphic analysis by Emerson. Figures 13-6 and 13-7 from his Section 13E are reproduced in Figures 4 and 5. His Figure 13-6 should be compared to Figure 1. His figure has retained the terms of model specification, but it has lost the detail of the factorial structure. It has combined the residual information from all the cells into a single column, and therefore lost the critical information that location in the material direction and spread are correlated. An attempt at recovering that information is in Figure 13-7, but the fitted values are not identified with their factorial labels. It is possible to see that the spread increases as fitted value increases, but the connection with the underlying levels of the factors is not visible.

By contrast, Emerson's earlier Section 13D presents an easily followed analysis and set of graphs for one-way models. The extension to the multiple factor models in his Section 13E is very hard to follow. The extension to multiple factor models by boxplot matrices presented here is more powerful. Figure 1 displays the model structure by the rows and columns of the boxplot matrix. It shows the effect of the factors, particularly Material, on predicted values and residuals as both level and variability increase with Material. It shows that, for this example, stabilization of within-cell variances is very important and that reduction of interaction is secondary.

Figure 5 shows a classical scatterplot (in the original scale) of within-cell residuals versus cell means. It too shows the non-homogeneity of variance in the Bekk data. But it also hides the two-factor structure (unless we use a complicated scheme of indicating the factor values by choice of plotting symbol). It makes more sense to compare this plot to the spread versus level plot in Figure 2 than to the data plot Figure 1, but it is in the wrong scale to be of use in calculating the best power for use in transformation.

3. BOXPLOT MATRICES AND THE ANOVA TABLE

The arithmetic for analysis of variance of balanced data is often presented as a summation technique over rows and columns of the cross-classification of the data. Table 2 shows the details for the two-way classification by Lab and Material.

An alternate formulation is as a regression over a set of dummy variables defined by the levels of the treatment factors. Table 3 indicates the details for the regression of $Y = \log(\mathsf{data})$ against an X matrix consisting of the set of dummy variables:

Effect	Dummy Variables			
Grand Mean	**1**			
Lab	L_1	L_2	L_3	
Material	M_1	M_2	M_3	M_4
Lab:Material	L_1M_1	L_2M_1	L_3M_1	
	L_1M_2	L_2M_2	L_3M_2	
	L_1M_3	L_2M_3	L_3M_3	
	L_1M_4	L_2M_4	L_3M_4	

defined by the Lab and Material factors. The dummy variables illustrated are based on the Helmert contrasts, the default contrasts used by the aov function in *S*.

	contrasts(bekk$lab)		
	L1	L2	L3
1	−1	−1	−1
2	1	−1	−1
3	0	2	−1
4	0	0	3

	contrasts(bekk$material)			
	M1	M2	M3	M4
1	−1	−1	−1	−1
2	1	−1	−1	−1
3	0	2	−1	−1
4	0	0	3	−1
5	0	0	0	4

With an orthogonal set of dummy variables, we can construct an orthogonal partitioning of the form

$$Y \quad = \quad Y_\mu + Y_{\mathsf{L}} + Y_{\mathsf{M}} + Y_{\mathsf{LM}} + Y_{\mathsf{res}}$$

Table 2. Row and column analysis of smoothness of paper in $\log_{10}$ units. The table illustrates that

$$\widehat{Y}_{lm} = \overline{Y} + (Y_{l.} - \overline{Y}) + (Y_{.m} - \overline{Y}) + (\widehat{Y}_{lm} - Y_{l.} - Y_{.m} + \overline{Y})$$

where

Grand Mean	$\overline{Y}$	$=$	$\text{avg}_{lm}\ \widehat{Y}_{lm}$
Cell Means	$\widehat{Y}_{lm}$	$=$	$\text{avg}_{k}\ Y_{lmk}$
Lab Means	$Y_{l.}$	$=$	$\text{avg}_{m}\ \widehat{Y}_{lm}$
Material Means	$Y_{.m}$	$=$	$\text{avg}_{l}\ \widehat{Y}_{lm}$
Lab effect	$Y_{l.} - \overline{Y}$		
Material effect	$Y_{.m} - \overline{Y}$		
Lab:Material effect	$\widehat{Y}_{lm} - Y_{l.} - Y_{.m} + \overline{Y}$		

$\widehat{Y}_{lm}$

	Material				
Lab	1	2	3	4	5
1	0.8281	1.0906	1.6573	2.1873	2.1519
2	0.7689	1.0636	1.6193	2.2212	2.3104
3	0.7683	1.0496	1.6040	2.1899	2.2176
4	0.6750	0.9925	1.5695	2.1169	2.1713

	$\widehat{Y}_{lm} - Y_{l.} - Y_{.m} + \overline{Y}$					
	Material					
Lab	1	2	3	4	5	$Y_{l.} - \overline{Y}$
1	0.04767	0.02112	0.02435	−0.01189	−0.08127	0.02039
2	−0.02523	−0.01946	−0.02724	0.00834	0.06360	0.03402
3	0.00499	−0.00269	−0.01169	0.00783	0.00155	0.00319
4	−0.02744	0.00103	0.01458	−0.00429	0.01611	−0.05761
$Y_{.m} - \overline{Y}$	−0.80259	−0.51359	0.04987	0.61616	0.65015	1.56266

Table 3. Regression analysis on dummy variables in $\log_{10}$ units.

The table below, illustrating

$$Y = Y_{\textsf{res}} + X\beta$$

was constructed by editing the output from the following S statements:

```
bekk.aov ← aov(log.y ∼ lab*material, proj=T, x=T, data=bekk)
b.rows ← c(1,2,3,4,5,6,40,80,120,160)
bekk[b.rows, "log.y", drop=F]
bekk.aov$x[b.rows,]
print(as.matrix(coef(bekk.aov)), digits=8)
proj(bekk.aov)[b.rows, "Residuals", drop=F]
```

The contrasts displayed in the X matrix are the Helmert contrasts.

			X																		
				L			M				LM										
Obs	Y	Y_{res}	(Int)	1	2	3	1	2	3	4	11	21	31	12	22	32	13	23	33	14	24	34
1	0.81291	−0.01523	1	−1	−1	−1	−1	−1	−1	−1	1	1	1	1	1	1	1	1	1	1	1	1
2	1.11059	0.02000	1	−1	−1	−1	1	−1	−1	−1	−1	−1	−1	1	1	1	1	1	1	1	1	1
3	1.58771	−0.06958	1	−1	−1	−1	0	2	−1	−1	0	0	0	−2	−2	−2	1	1	1	1	1	1
4	2.22089	0.03357	1	−1	−1	−1	0	0	3	−1	0	0	0	0	0	0	−3	−3	−3	1	1	1
5	2.10003	−0.05191	1	−1	−1	−1	0	0	0	4	0	0	0	0	0	0	0	0	0	−4	−4	−4
6	0.81291	−0.01523	1	−1	−1	−1	−1	−1	−1	−1	1	1	1	1	1	1	1	1	1	1	1	1
:																						
40	2.25527	0.10333	1	−1	−1	−1	0	0	0	4	0	0	0	0	0	0	0	0	0	−4	−4	−4
:																						
80	2.26928	−0.04116	1	1	−1	−1	0	0	0	4	0	0	0	0	0	0	0	0	0	4	−4	−4
:																						
120	2.29667	0.07911	1	0	2	−1	0	0	0	4	0	0	0	0	0	0	0	0	0	0	8	−4
:																						
160	2.29842	0.12710	1	0	0	3	0	0	0	4	0	0	0	0	0	0	0	0	0	0	0	12

β
1.56266
0.00681
−0.00801
−0.01921
0.14450
0.23599
0.25957
0.16254
0.00808
0.00045
0.00475
0.00086
−0.00060
0.00309
0.00941
0.00136
−0.00003
0.01811
0.00087
0.00134

The numerical values of this partitioning for the Bekk smoothness data are in Table 4. Each of the 5 columns on the right side of Table 4 is the projection of the dependent variable Y onto a subspace defined by the corresponding dummy variables. For example, the Lab.effect, Y_{L} with 3 degrees of freedom, is the projection of Y onto the three dummy variables defining the Lab effect:

$$Y_{\mathsf{L}} = L_1\widehat{\beta}_{L_1} + L_2\widehat{\beta}_{L_2} + L_3\widehat{\beta}_{L_3}$$

The corresponding sums of squares are additive:

$$Y'Y \quad = \quad Y'_\mu Y_\mu + Y'_{\mathsf{L}} Y_{\mathsf{L}} + Y'_{\mathsf{M}} Y_{\mathsf{M}} + Y'_{\mathsf{LM}} Y_{\mathsf{LM}} + Y'_{\mathsf{res}} Y_{\mathsf{res}}$$

These are the familiar "Sums of Squares" from the ANOVA table.

3.1 Calculations

The partitioning into orthogonal projections can be calculated, for example, by the ANOVA function aov(formula, data, projections=T, ...) in the August 1991 version of S (Chambers, Freeny, and Heiberger (1991)). For the Bekk data, the usage is:

```
bekk.aov ← aov(log.y ∼ lab*material, data=bekk, projections=T)

boxmat(proj(bekk.aov)[,"lab"] ∼ lab+material, data=bekk)
```

Each column in the projection matrix, hence each boxplot matrix in a series displaying the lines of an ANOVA table, is based on an orthogonal decomposition corresponding to a forward sequential regression analysis on a set of dummy variables defined by the factor structure. In many ANOVA situations, balanced designs in particular, the dummy variables are orthogonal. In other cases (with covariates, partial confounding, or unbalanced cell structures) they are not. An orthogonal decomposition is always possible, although it

Table 4. Projection onto dummy variables in $\log_{10}$ units.

The table below, illustrating

$$Y \quad = \quad Y_\mu + Y_{\mathsf{L}} + Y_{\mathsf{M}} + Y_{\mathsf{LM}} + Y_{\mathsf{res}}$$

was constructed by editing the output from the following *S* statements:

```
b.rows ← c(1,2,3,4,5,6,40,80,120,160)
bekk[b.rows, "log.y", drop=F]
proj(bekk.aov)[b.rows,]
```

Obs	Y	Y_μ	Y_{L}	Y_{M}	Y_{LM}	Y_{res}
1	0.81291	1.56266	0.02039	−0.80259	0.04768	−0.01523
2	1.11059	1.56266	0.02039	−0.51359	0.02113	0.02000
3	1.58771	1.56266	0.02039	0.04987	0.02436	−0.06958
4	2.22089	1.56266	0.02039	0.61616	−0.01189	0.03357
5	2.10003	1.56266	0.02039	0.65015	−0.08127	−0.05191
6	0.81291	1.56266	0.02039	−0.80259	0.04768	−0.01523
:						
40	2.25527	1.56266	0.02039	0.65015	−0.08127	0.10333
:						
80	2.26928	1.56266	0.03402	0.65015	0.06360	−0.04116
:						
120	2.29667	1.56266	0.00319	0.65015	0.00155	0.07911
:						
160	2.29842	1.56266	−0.05761	0.65015	0.01612	0.12710

is not necessarily unique. All the usual discussions of the correct order in which to introduce terms into the model apply.

3.2 Graphic Displays

The last 5 columns of Table 4 are an $(n \times p) = (160 \times 5)$ data matrix. The familiar geometry of least squares is based on thinking of each row of an $(n \times p)$ data matrix as a p-dimensional point in p-space. This leads to plots similar to the $(n \times 2)$ plot in Figure 5 with the residuals Y_{res} on the ordinate and a continuous variable (in this case the fitted value $\widehat{Y} = Y_\mu + Y_{\mathsf{L}} + Y_{\mathsf{M}} + Y_{\mathsf{LM}}$) on the abscissa. This type of plot does not display all columns, nor does it show the orthogonality relations among them.

The "natural" geometry associated with least squares techniques is the n-dimensional dual geometry, in which each column of the data matrix is viewed as an n-dimensional point in n-space. In this setting, an orthogonal projection is exactly a dropping of a perpendicular line from the point corresponding to the response variable to a hyperplane defined by the predictor variables. Figure 6(a) shows the geometry for the projection of $Y - Y_\mu$ onto $\widehat{Y} - Y_\mu$ in the space spanned by the dummy variables for the treatment factors. Figure 6(b) expands the projection of $\widehat{Y} - Y_\mu$ onto the projections Y_{L}, Y_{M}, and Y_{LM} associated with the sets of dummy variables for the main effects Lab and Material and their interaction. A fairly complete discussion of the n-space geometry of ANOVA and regression is given in Heiberger (1989).

Although 3-dimensional statisticians are quite comfortable with many characteristics of n-space, visualization is not one of them. Visualization is even more difficult for non-statisticians. The boxplot matrices in Figure 3 give a more easily visualized and interpreted picture of n-space geometry than do Figures 6(a) and 6(b). Each subfigure in Figure 3 plots one column of Table 4, hence one line of the ANOVA table. The $n = 160$ numbers in Y_{L} (with only 4 distinct values), eight (all alike) per cell for each of the $4 \times 5 = 20$ cells, are plotted in Figure 3(d). The sum of squares of these 160 numbers, $Y_{\mathsf{L}}'Y_{\mathsf{L}}$, is the sum of squares for Lab.

Because the projections onto columns are defined in terms of the classification factors of the design, the level and spread differences among the various boxplots in a matrix reflect the effects of the defining factors. We are able to use our familiarity with classification factors to understand some characteristics of the n-dimensional geometry.

Systematic changes in level or spread among the individual boxplots in a boxplot matrix can be visually correlated with changes in level of the factors defining the matrix. The formal assessment of the differences between means in a boxplot matrix is described in terms of the usual SED (standard error of differences) calculated from the residual standard deviation and the number of levels of each of the factors. Thus, a series of boxplot matrices provides a visualization of the multiple comparisons problem applied to multiple factor situations.

For simple two-way designs, and for main effects in more complex designs, some of the boxplot matrices may be redundant (the columns of 3(d) and the rows of 3(e) are identical), or superfluous (each individual boxplot is based on a five-number summary and in subfigures 3(d), 3(e), and 3(f) all five numbers reduce to a single value). In such cases, it may suffice to display a single boxplot matrix for the combined treatment effect (subfigure 3(b)).

The boxplot matrices in Figure 3 are individually scaled. An alternate scaling, in which all boxplot matrices are scaled in the original units, might also be appropriate. The alternate scaling is natural in the sense that all columns are projections of the same data onto different subsets of the dummy variables. A third appropriate scaling for a set of boxplot matrices would be to make a common distance (say, one inch) equal to one standard error of differences within each boxplot matrix in a series.

4. ANALYSIS OF COVARIANCE IN A BLOCKED DESIGN

Consider, for example, the classic eelworm data [Cochran and Cox (1957, Ch. 3) and Heiberger (1989, Ch. 18.5)]. In 1935 an experiment was conducted at Rothamsted Experimental Station to measure the effectiveness of four soil fumigants in keeping down the number of eelworms in the soil. Each fumigant was tested in both a single and a double dose, thus providing eight active treatment combinations. A control (no fumigant) was the ninth treatment. The response variable second count is the number of eelworm cysts per 400 grams of soil after harvest. The covariate first count is the naturally occurring infestation prior to any treatment, in this case the number of eelworm cysts counted before the introduction of the treatments. There are two treatment factors (fumigant and dose) and a blocking factor (block). The control treatment is observed four times in each block, the active treatments are observed once in each block. Some cells implied by the crossing of the treatment factors are empty (the control fumigant is only at dose level 0, the active treatments are only at dose levels 1 and 2).

The boxplot matrix of the dependent variable second count is displayed in Figure 7(a). The standard analysis of covariance for the eelworm data from Cochran and Cox Table 3.9 and Heiberger Table 18.5.18 is in Table 5. In Figure 7 separate boxplot matrices are displayed for each individual line of

Table 5. Analysis of covariance of Eelworm response variable second count. The analysis may be specified by the *S* statement:

```
cc46.aov ← aov(second.count ~ block + first.count + (ctl.vs.trt/dose/fumigant),
   proj=T, data=cc46)
```

Source	Df	Sum of Sq	Mean Sq	F Value	Pr(F)
block	3	289427	96476	13.5280	0.000005
first.count	1	215343	215343	30.1958	0.000004
ctl.vs.trt	1	105911	105911	14.8510	0.000476
dose %in% ctl.vs.trt	1	2816	2816	0.3949	0.533835
fumigant %in% (ctl.vs.trt/dose)	6	128463	21411	3.0022	0.017918
Residuals	35	249605	7132		

the ANOVA table. These plots enable us to visualize both the contrasts and variability associated with the effects.

The eelworm data was collected from a blocked design. The first step in the analysis is the removal of the block effect in Figure 7(b), leaving the second count adjusted for blocks in Figure 7(c). Note that the within-cell variability (as measured by the hinge-spread or inter-quartile distance of the individual boxplots) is noticeably reduced. Almost 30% of the sum of squares (13% of the standard deviation) has been explained by only 6% of the degrees of freedom.

The next step is to look at the effect of the covariate first count. Figure 8 shows the scatterplot of $y =$ second count against $x =$ first count, both adjusted for blocks. The $\widehat{y}$ values from the regression line in Figure 8, the effect of the covariate, are plotted in Figure 7(d). When we remove the effect of the covariate from Figure 7(c) to get Figure 7(e), we see that the covariate has explained much of the remaining variability.

The final step is to remove the treatment effect in Figure 7(f) leaving the residual in Figure 7(g). Comparison of these last two plots provides the F test for treatments adjusted for blocks and the covariate. Note that values in each of the cells of the treatment boxplot matrix Figure 7(f) are not constant. The within cell variability is a consequence of the removal of the effect of the covariate.

Comparisons among the treatments are best made by expanding the scale of Figure 7(f) into Figure 9(a) and transposing it in Figure 9(b). The comparisons within single and double doses from their table on p. 68 are visible in Figure 9(a). The curvature effects in Cochran and Cox's Table 3.5 are visible in Figure 9(b).

Standard plots of interactions, as in Figures 9(c) and 9(d), force an asymmetry in the interpretation of the factors. Plotting dose along the abscissa implicitly suggests that dose is a discretization of a continuous variable, and plotting distinct lines for the levels of factor fumigant implicitly suggests that it is categorical. The asymmetry in Figure 9(d) accurately reflects the factor definitions. The transpose of this plot in Figure 9(c), with fumigant along the abscissa, inaccurately suggests that fumigant is a continuous variable.

The boxplot matrix imposes no such asymmetry. Both row and columns factors are treated alike as discrete. Although comparisons among rows of a boxplot matrix are easier to see than comparisons among columns, transposition does not challenge the implicit discreteness of the factors.

5. EXTENSIONS AND USAGE CONSIDERATIONS

5.1 Higher-Order Designs

The boxplot matrix can easily be extended to an arbitrary number of factors. Three or four factors are relatively easy to understand. Larger numbers of factors are more difficult to interpret.

It is fairly easy to rearrange data manually or automatically into sets of two-way classifications for plotting. (In *S* the ":" operator is used to construct a new factor jointly indexed by two existing factors. Thus the formula `a:b * c` specifies an $n_a n_b \times n_c$ classification which is distinct from the `a * b:c` specification of an $n_a \times n_b n_c$ classification.) Several alternative mappings, in which multiple factors are combined in different ways and in different orders, should be investigated before determining the optimal displays. Higher-order plots may sometimes be better presented as entirely separate two-factor plots.

Generally a boxplot matrix is easier to look at if there are fewer *rows* than *columns*. Usually horizontal comparisons across boxes are easier than vertical ones. Thus it is important to be able to transpose a plot and retain both the customized labeling information and the correct linkages to other plots.

5.2 Series of Plots

A series of plots, for example of the lines of an ANOVA table, should have the same customization of labels and margins. It is easiest to write a cover function tailored to the data set and then call the cover function for each of the individual plots. In the object-oriented setting of XLISP-STAT, the first plot can be constructed incrementally. Later plots can then query the first one to determine its customization.

Initially, plots in a series should have the same scaling. After the within-cell variability of boxplots has been reduced by removal of blocking and covariate effects, it is often helpful to display further analysis of treatment factors on an expanded scale.

5.3 Display of Matrices

The structure of the boxplot matrix can be used to display the values of a numerical matrix. This is often helpful when viewing transition matrices or distribution matrices. Figure 10 shows the progression of a subset of an entry cohort of university Freshmen through classifications (Withdraw, Freshman, Sophomore, Junior, Senior, High Senior, Graduate) over a six-year time period. Each column of the matrix is a histogram of the observed class distribution at the end of the stated number of years. Further discussion of this data appears in a separate paper (Heiberger, 1992).

There are two steps to the construction of this display. First, each number n_{ij} in the data matrix is represented by a boxplot of the pair of numbers $(0, n_{ij})$ The pairing is done in the *S* implementation by the function nwaylist.c0a. Negative values n_{ij} are acceptable. Missing values n_{ij} are represented by bypassing the boxplot for the ij cell. Second, the median line that usually bisects each of the boxplots is suppressed. The *S* implementation does this by replacing the default value boxf=boxplot with the argument boxf=boxplotm. The function boxplotm intercepts the five number summary produced by the standard boxplot and changes the value of the third position, the place where the median is usually stored. In effect it replaces the five-number summary for each individual boxplot by a summary that has the first three numbers equal to $\min(0, n_{ij})$ and the last two numbers equal to $\max(0, n_{ij})$.

5.4 Alternate Five-Number Summaries

Because ANOVAs are computed using least squares techniques, we may want to see means as well as (or instead of) medians on the boxplots. Although it is easy to define a function that uses an alternate five-number summary and to force its use with the boxf argument, doing so loses the visualization of skewness afforded by the standard definition of boxplots based on the median and hinges (quartiles). Except for the matrix display example in Section 5.3, all boxplot matrices displayed here use the order statistics.

6. DEVELOPMENT OF THE BOXPLOT MATRIX

There were several key steps in the development of this program.

An earlier version of the boxmat function was developed in S with Anne E. Freeny. An example of its use appears in Freeny and Landwehr (1990). In that version of the boxmat function we used multiple calls to the S boxplot function, one call per row of the boxplot matrix. The function used the multiple frame per physical page option (par(mfrow=c(4,1))), and much of its detail revolved around minimizing the margins around each frame to give the appearance of a single plot.

The second level of development followed from the paradigm switch that came with learning XLISP-STAT. Although XLISP-STAT permits more than one graphics window to be on the terminal screen simultaneously, it (more specifically, the MS-WINDOWS version) does not give the user program control over placement of the windows on the screen. It therefore became imperative to place the entire boxplot matrix within a single frame. The technique, precalculating the vertical offset to force each row of the boxplot matrix into a different vertical piece of the window, was then applied back to the *S* function.

The next step, which works in XLISP-STAT but not yet in *S*, uses the feature of linked graphs. Multiple views of the same data set, for instance boxplot matrices corresponding to the lines of an ANOVA table, can be linked. It is possible, for example, to color a boxplot for a cell in the residual boxplot matrix and see the color propagate back to all other boxplot matrices in the series.

Once it was recognized that a plot of the numerical values in a matrix is similar to a boxplot matrix, it was easy to parameterize the call to the boxplot function.

The final step so far uses the model language of the August 1991 version of *S* (Chambers and Hastie, 1991). It is now possible to specify a boxplot matrix directly from the formula describing the plotted variable and the factor structure. The boxmat function then automatically constructs the two-way list of data vectors within each cell.

7. PROGRAM AVAILABILITY

The specific *S* and XLISP-STAT functions we describe are available by electronic mail or *ftp* transfer from the statlib server maintained at Carnegie Mellon University. Anyone with internet access can send *e-mail* messages of the form:

```
echo "send boxmat from S" | mail statlib@lib.stat.cmu.edu
echo "send boxmat from XlispStat" | mail statlib@lib.stat.cmu.edu
```

A discussion of electronic libraries of computer code is available in Dongarra and Grosse (1987). Information on the statlib server is obtainable by *e-mail* with the following message:

```
echo "send index" | mail statlib@lib.stat.cmu.edu
```

Some of the power of the boxplot matrix can be obtained with a boxplot function designed to display one-way parallel boxplots. We can construct a one-way factor from the two or more factors describing the data by using standard techniques for indexing the cells of the data. For example, we would construct an index factor for the Bekk data by a statement of the form:

```
index = 10*lab + material
```

and then construct the parallel boxplots based on the factor index.

APPENDIX: CONSTRUCTION OF BOXPLOT MATRICES

While the interpretation of a matrix of boxplots is relatively straightforward, construction of the matrix is not. In this section we describe the issues that must be addressed.

We assume the existence of a function that can construct a set of parallel boxplots along a single classification factor. Parallel boxplots are currently available in many statistics packages and are often used as a graphical display in one-way ANOVA problems. Emerson and Strenio (1983) give several examples with discussion.

The boxplot was originally defined by Tukey (1970, 1977). A detailed discussion on the programming of boxplots appears in Velleman and Hoaglin (1981). A comparison of several common implementations, with discussion on the effects of alternate definitions of the quartiles, is given by Frigge, Hoaglin, and Iglewicz (1989). Additional features have been added to the basic boxplot. McGill, Tukey, and Larsen (1978) suggested varying the width of the individual boxes as a function of sample size and adding notches as an indicator of rough significance of differences between medians. Further modifications in the visual appearance of the boxplots have been suggested by Tukey (1990).

We list the additional considerations needed in the extension to a matrix of boxplots. Our presentation is general, but we also provide implementations in both *S* (Becker, Chambers, and Wilks 1988) and XLISP-STAT (Tierney 1990) that incorporate these ideas. The functions can be requested by *e-mail* or *ftp* as described in Section 7.

Our list of specific details that must be attended to for a single two-factor matrix of boxplots is:

1. Row and column alignment. The requirement for column alignment makes the use of variable width boxes difficult. In *S* the use of the varwidth parameter to the boxplot and bxp functions to make the widths proportional to the square root of the number of observations also changes the horizontal distance between box centers. In XLISP-STAT the :add-boxplot message allows independent control of box widths and box centers.

2. Empty cells in the cross-classification. In *S* each row of the boxplot matrix is processed by a single call to boxplot. Zero-length within-cell vectors are converted to NA for correct handling by boxplot. In the XLISP-STAT function, each cell is plotted by a separate call to the :add-boxplot message. The call is bypassed for empty cells.

3. Vertical spacing. In both languages the system-provided boxplot routine was designed for a single row of parallel boxplots. In the initial design of the boxplot matrix we used program control (the mfrow parameter and margin control in *S*) to place the graphs from several independent calls to the boxplot function on same physical page. In the redesign, first in XLISP-STAT and then in *S*, we precalculated the vertical offset of each row of boxplots and placed all rows on the same graph. In *S* this required the new=T parameter.

4. Vertical scaling. All rows of a single boxplot matrix must be scaled alike. Different boxplot matrices in a series may be scaled differently.

5. Column-factor axes. In *S* we suppressed the automatically produced boxes, x- and y-axes, and axis labels. In XLISP-STAT we suppressed the automatically produced y-axis. We replaced them with our own axes for the levels of the column factor, with a new set of labels for the row factor, and with a common scale for the vertical axis within each row of boxes. In both we arranged for horizontal lines separating the rows (and when there are more than two factors, vertical lines separating the columns) of the boxplot matrix.

6. Graphics devices. Different devices have different labeling capabilities. We found it necessary to tune the graphics parameters for each device we used. The left margin, the area where the row factor labels and the y-axis values appear, often needs to be adjusted for each data set.

7. Linking plots. In the XLISP-STAT version a boxplot matrix is an object that inherits from the scatterplot matrix and, in addition, has several other methods attached to it. The boxmat function in XLISP-STAT

is designed to link plots, permitting brushing or coloring of one view of a cell to be reflected by coloring the corresponding cell of linked plots. The correspondence holds for both the original and transposed orientation of all boxplot matrices associated with an ANOVA table.

8. Multiple plots. Many plots can be simultaneously visible in XLISP-STAT. Multiple boxplot matrices can therefore be displayed on the same screen for immediate interactive visual comparison. *S* has only one graphics window visible at a time (August 91 version). The mfrow parameter is used to make several boxplot matrices appear on the same page or screen.

9. Alternate five-number summaries. Boxplot matrices of the numerical values of data matrices look better when the median line is suppressed. In *S*, setting the argument boxf=boxplotm replaces the standard boxplot function with a variant that intercepts the five-number summary before letting bxp continue with the actual plot. In XLISP-STAT, the :add-boxplot method is replaced with a much simpler alternate method that just draws rectangles.

NOTE

The calculation of projections using *S* in Section 3 assumes that the corrected proj.default has been installed. I posted the correction to S-news on June 11, 1992. It is also included in the statlib distribution for the boxplot matrix paper.

ACKNOWLEDGMENT

I would like to thank the anonymous referees and associate editor whose comments have strengthed the paper.

REFERENCES

Becker, R. A., J. M. Chambers, and A. R. Wilks (1988), *The New S Language: A Programming Environment for Data Analysis and Graphics*, Wadsworth, Monterey, CA.

Chambers, J. M., A. E. Freeny, R. M. Heiberger (1991), "Analysis of Variance; Designed Experiments." Chapter 5 (pp. 145–193) in *Statistical Models in S*, edited by J. M. Chambers and T. J. Hastie, Wadsworth, Monterey, CA.

Chambers, J. M., and T. J. Hastie, editors (1991), *Statistical Models in S*, Wadsworth, Monterey, CA.

Cochran, W. G., and G. M. Cox (1957). Experimental Designs, Wiley, New York.

Dongarra, J. J., and E. Grosse (1987), "Distribution of Mathematical Software by Electronic Mail." *Communications of the ACM*, **30** (5), 403–407.

Emerson, J. D. (1991), "Introduction to Transformation." In D. C. Hoaglin, F. Mosteller, and J. W. Tukey (eds.). *Fundamentals of Exploratory Data Analysis*, pp. 365–400, Wiley, New York.

Emerson, J. D., and J. Strenio (1983), "Boxplots and Batch Comparison." In D. C. Hoaglin, F. Mosteller, and J. W. Tukey (eds.) *Understanding Robust and Exploratory Data Analysis*, pp. 58–96, Wiley, New York.

Freeny, Anne E., and James M. Landwehr (1990), "Displays for Data from Large Designed Experiments." *Computing Science and Statistics: Proceedings of the 22nd Symposium on the Interface*, pp. 117–126, Springer-Verlag, New York.

Frigge, M., D. C. Hoaglin, and B. Iglewicz (1989), "Some Implementations of the Boxplot." *The American Statistician*, **43**, 50–54.

Heiberger, Richard M. (1989). *Computation for the Analysis of Designed Experiments*, Wiley, New York.

Heiberger, Richard M. (1992). "Survival Analysis of Cohort Progression with Applications using University Student Retention Data," submitted to *Journal of Educational Statistics.*

Mandel, J. (1964). *The Statistical Analysis of Experimental Data*, Wiley, New York.

McGill, R., J. W. Tukey, and W. A. Larsen (1978), "Variations of Box Plots." *The American Statistician*, **32**, 12–16.

Tierney, Luke (1990). *LISP-STAT: An Object-Oriented Environment for Statistical Computing and Dynamic Graphics*, Wiley, New York.

Tukey, J. W. (1970). *Exploratory Data Analysis—Limited Preliminary Edition*, Addison-Wesley, Reading, MA.

Tukey, J. W. (1977). *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

Tukey, J. W. (1990). "Data-Based Graphics: Visual Display in the Decades to Come," *Statistical Science*, **5**, 3, 327–339.

Velleman, P. F., and D. C. Hoaglin (1981). *Applications, Basics, and Computing of Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

Figure 1. Measurements on the smoothness of paper using the Bekk method (Mandel 1964). Measurements on five types of paper observed at four different laboratories.

Figure 2. Spread-vs-level plot of $\log_{10}(\textsf{cell standard deviation})$ against $\log_{10}(\textsf{cell mean})$ of data in Figure 1. The slope of 1.1318 suggest a transformation to the power $1 - 1.1318 = -0.1318$. We round this to the power 0 and use the logarithm.

Figure 3. Boxplot matrix analysis of the logarithms of the paper smoothness data in Figure 1.

3(a). log paper smoothness.

3(b). Treatment effects of log paper smoothness.

3(c). Within-cell variability of log(data).

3(d). Laboratory effect.

3(e). Material effect.

3(f). Laboratory:Material interaction.

Figure 4. Figure 13.6 from Emerson (1991).

Figure 5. Figure 13.7 from Emerson (1991).

Figure 6. Dual space plots of projections of the 160-vector $Y - \overline{Y}$ onto subspaces spanned by the dummy variables for the treatment factors. The plots are scaled to show each vector's length equal to the square root of its sum of squares in the ANOVA Table 1.

6(a). Dual space plot of $Y - \overline{Y}$ projected onto the treatment space $\widehat{Y} - \overline{Y} = Y_{\mathsf{L}} + Y_{\mathsf{M}} + Y_{\mathsf{LM}}$ and residual space Y_{res}.

6(b). Dual space plot of the projection of $\widehat{Y} - \overline{Y}$ onto the sets of dummy variables for the main effects Lab and Material and their interaction, Y_{L}, Y_{M}, and Y_{LM}.

Figure 7. Analysis of covariance of the eelworm data (Cochran and Cox 1957, Ch. 3).

7(a). Response variable second count.

7(b). Block effect block.

7(c). second count adjusted for block.

7(d). Effect of the covariate first count adjusted for block.

7(e). second count adjusted for block and the covariate.

7(f). Treatment effect dose*fumigant adjusted for block and the covariate.

7(g). Residual after block and treatments and covariate.

Figure 8. Scatterplot of $n = 48$ points in 2-space: the response variable second count is plotted against the covariate first count, with both variables adjusted for the block effect.

Figure 9. Treatment effect dose*fumigant adjusted for block and the covariate.

9(a). Boxplot matrix showing comparisons within single and double doses. Figure 7(f) in an expanded scale.

9(b). Transposed boxplot matrix showing curvature as a function of dose within each fumigant.

9(c). Line plot of treatment effects with inaccurate suggestion that fumigant is a continuous variable.

9(d). Line plot of treatment effects showing curvature as a function of dose within each fumigant.

Figure 10. Progression of a subset of an entry cohort of university Freshmen through classifications (Withdraw, Freshman, Sophomore, Junior, Senior, High Senior, Graduate) over a six-year time period. The height of the ijth box is proportional to the number of entering students who have attained the ith class at the completion of the jth year. Each column of the boxplot matrix shows the distribution of students as of the end of the stated number of years. The sum of the box heights in each column is exactly 1. The drop in the Withdraw proportion after year 6 is correct. Students frequently take a semester or year off before reenrolling and completing their degree.