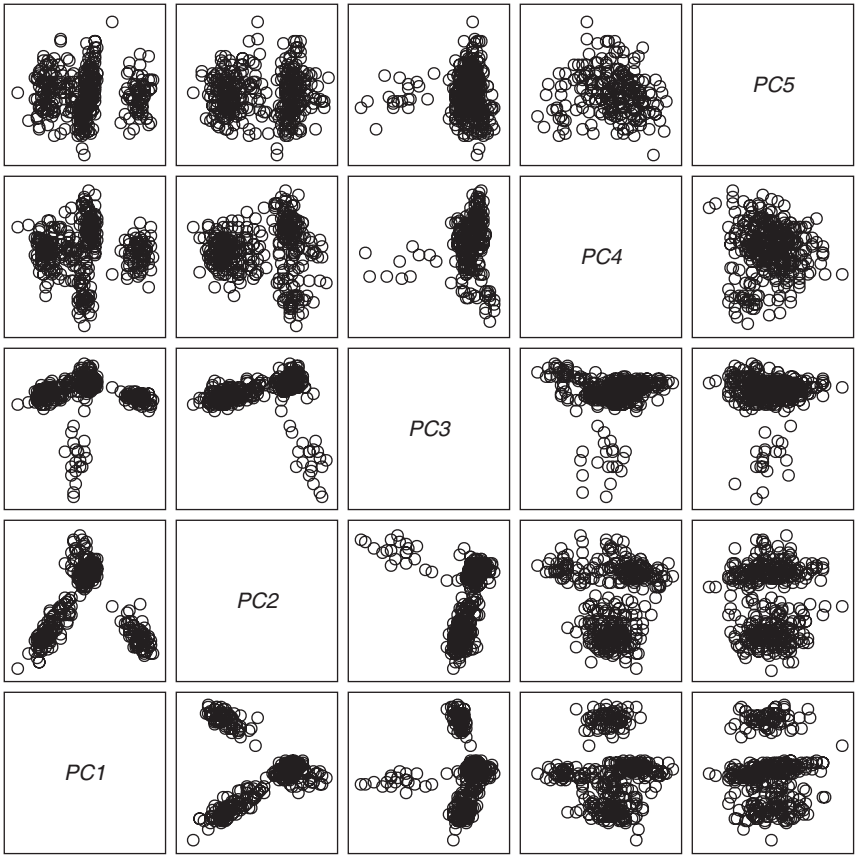
# 2 Examples



# 2 Examples

In this chapter we present three examples of visual statistics where dynamic interactive graphics are central in the discovery of structure in data. The first example shows how a spinning cloud of points can reveal structure in data that is very difficult to see using static views of the same point cloud. The second example uses data concerning skin disease to illustrate how one uses dynamic interactive graphics to develop a classification scheme. The third example uses data about sexual behavior and its effects on marriage and divorce.

## 2.1 Random Numbers

This example shows how a specific dynamic graphic—a plot showing a spinning cloud of points—can reveal structure in data that is very difficult to see using static views of the same point cloud. The example uses numbers generated by Randu, a random number generator widely used during the 1960s and 1970s. Marsaglia (1968) showed that Randu, and other members of the linear congruential family of random number generators, yield numbers that are nonrandom in a special way which can be revealed by dynamic graphics.

Randu is supposed to generate uniformly distributed random numbers between 0 and 1 (inclusive). These numbers should be generated so that all equal-width subintervals within the [0,1] interval are equally likely (uniform), and so that the next number to be generated is unpredictable from the numbers already generated (random).

We use one-, two, and three-dimensional graphics to look for structure or pattern in the numbers. Since the numbers are random, there should be no structure, so seeing structure implies nonrandomness. We will use graphical techniques that are presumably unfamiliar—sequence plots, lag plots, jittered dot plots, spinplots, and so on. We begin by showing how these plots look with normally distributed data, since we assume that such data are more familiar than uniform data.

**Normal distribution.** A visualization of 3000 univariate normal random numbers is shown in Figure 2.1. The visualization involves four graphics (all are discussed more extensively in Chapter 6). The upper-left graphic is a dot plot. The dots are located vertically according to their generated value, and horizontally according to their frequency within a narrow range of generated values. Since the number of dots within a narrow range of generated values reflects density, the shape of the distribution, as represented by the number of dots, should reflect the population distribution, which is normal. We see that it does.

The upper-right graphic is a shadowgram. It is based on averaging many histograms together in a way that determines the density of the distribution for each pixel in the graph. The graph is then rendered so that density is shown by the darkness of the shade of gray—the higher the density the darker the shade. Again, the plot should look like the parent population (i.e., be normal in shape), which it does.

The lower-left graphic is a lag plot. This plot shows each datum plotted against the datum that was generated previously (a lag of 1). This gives us a two-dimensional

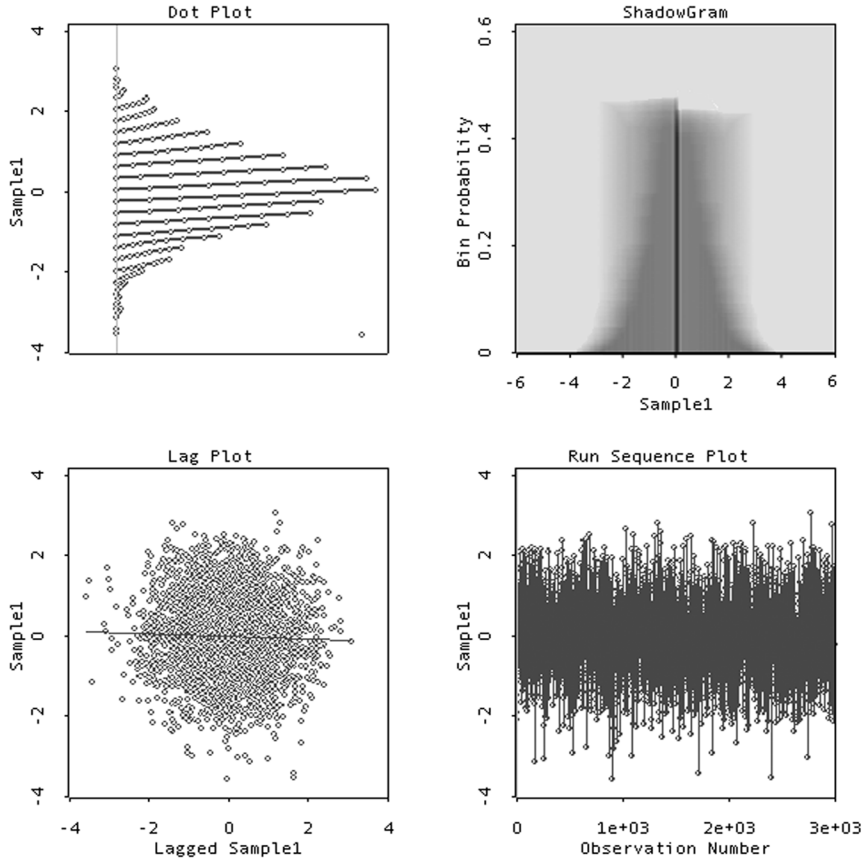



**Figure 2.1 Four graphs of 3000 normally distributed random numbers.**

plot where the first dimension is the values generated at time $t$-1 and the second dimension is for the values generated at time $t$. For normally distributed values, each dimension should be normal and the two dimensions should be uncorrelated. Also, each dimension should have the same mean and variance (0 and 1 for standard normal values). When plotted against each other, the plot should be bivariate normal, which will look circular in shape, with the density highest in the middle and tapering off normally in all directions. The figure looks as it should.

Since the values are random, they should be unrelated to the order in which they were generated. The sequence plot of values versus generation order should show no structure. This plot is shown in the lower right of Figure 2.1. It has no discernible structure.

**Uniform Distribution.** We turn now to the data generated by the faulty RANDU random number generator. In Figure 2.2 we see the way that 3000 supposedly random

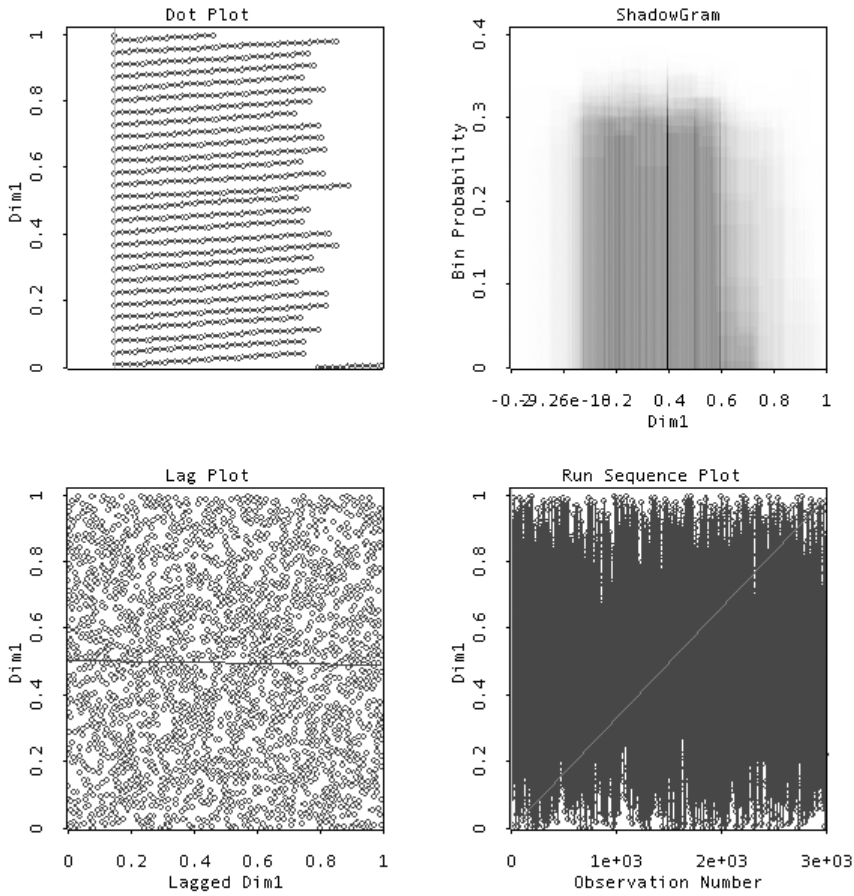



**Figure 2.2 Four graphs of a uniformly distributed random variable.**

numbers generated by RANDU look with the same graphics as those used for the normally distributed numbers in Figure 2.1. The upper-left graphic in Figure 2.2 is the dot plot for the values generated by RANDU. Since the distribution should be everywhere equally dense, the width of the plot should be the same for all values generated. This appears to be the case.

The upper-right graphic shows an estimate, from our sample, of our sample's population distribution. If our sample is good, the estimate based on it ought to be flat, which it seems to be (the right side drop-off is an artifact of the estimation method). The autocorrelation (correlation of values generated with themselves with a specified lag) should be zero, as it is in the lower-left figure for a lag of 1 (this holds up when we cycle through various lags). Finally, the sequence plot (lower right of Figure 2.2) should show no pattern, which seems to be the case.

So far, so good! The one-dimensional views of the random numbers look as they should if the values generated by the random number generator are indeed random uniform numbers. (Find out more about these and other one-dimensional views of data in Chapter 6.) In addition to the one-dimensional views, we can look at two-dimensional, three-dimensional, and even higher-dimensional views. We do that now.

For a two-dimensional view, we use the 3000 supposedly random uniform numbers to represent 1500 observations of two variables. The main method for assessing the relationship between two variables is the well-known scatterplot, which is shown for our data in Figure 2.3. If the axes of the scatterplot are two random uniform variables, we should obtain a scatterplot with points scattered with equal density everywhere inside a square defined by the interval [0,1] on each axis. There should be no discernible pattern. Looking at Figure 2.3, our numbers, once again, look random.

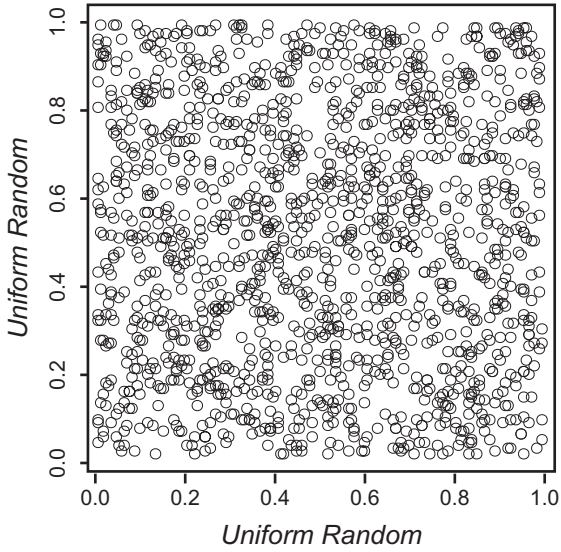



**Figure 2.3 The 3000 uniform numbers shown as 1500 2D points.**

We can do essentially the same thing in three dimensions that we just did in two dimensions: Arrange our 3000 supposedly uniform random numbers to represent the coordinates of 1000 points in three dimensions. We can then form a spinning three-dimensional scatterplot and look at it for structure. We have done this in Figure 2.4 for two very slightly different rotations of the space.

On the left we see that the space looks like the featureless space it should be if the data are truly random. However, if we watch the cloud of points slowly rotate, or (better yet) if we use the rotation tool to actively rotate the cloud of points, we will eventually see a view of the space that reveals the nonrandom structure lurking within our points. We see the structure in the right space in Figure 2.4. What we see is that the points are arranged in a set of parallel planes within the point cloud. We cannot see this in any univariate or bivariate plot, since the planes are not lined up with the dimensions of the space.

As pointed out by Wainer & Velleman (2001), the effect is like what you see as you drive past a field of corn. Most of the time you just see an apparently random arrangement of cornstalks, but occasionally the cornstalks line up and reveal their nonrandom structure.

Furthermore, if we did not have a dynamic graphic to look at the trivariate space, we wouldn't see the structure either. It took a dynamic graphic, in this case a three dimensional spinnable space, to reveal the structure, and although it wasn't necessary, having the ability to interact with the space made it easier to find the structure.

Finally, note that the parallel plane structure is very well hidden in the space—a very small rotation can obscure the structure completely. For example, the two views in Figure 2.4 differ by only $3^{o}$. If you compare the positions of the boxes, you can see how small a difference this is.

If all you see is the left-hand view, you have no idea that there is hidden structure to be revealed with just a $3^{o}$ rotation. Then, when you see the right-hand figure, the

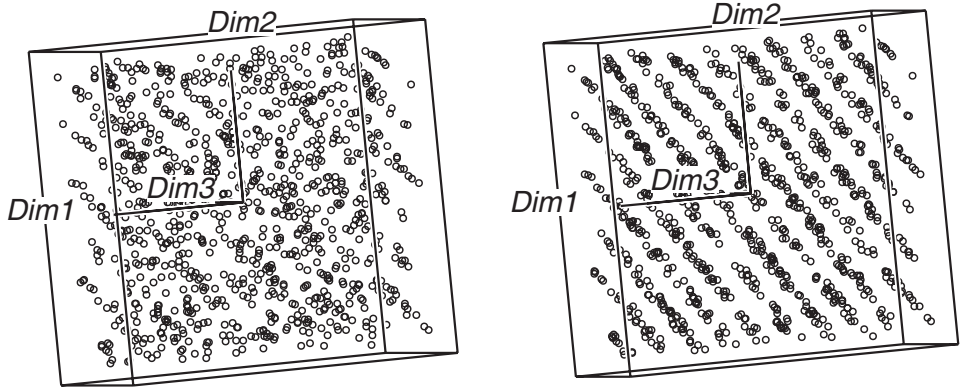



**Figure 2.4 Two views of the same 3000 uniform random numbers, now seen as 1000 points in three dimensions.**

structure is revealed, and the IOI test is satisfied. Of course, having now seen the right-hand figure, where the structure has clearly emerged, you can see it beginning to appear in the left-hand view, but you wouldn't see it if you didn't know what to look for.

This is, by the way, a nice example of the IOI test at work. When we see the space rotating, there is a bit of a "glitch" in the rotation; it has already "hit us between the eyes," if only for a moment. Then, if we stop the rotation and back it up somewhat, we see the structure and are amazed.

## 2.2 Medical Diagnosis

In this section we show dynamic interactive graphics being used to lay the groundwork needed to develop a medical diagnostic tool. We present a more complete version of this example in Chapter 8 where we use dynamic interactive graphics to develop the tool, and where we show how the tool would be used by the diagnostician to assist in making a diagnosis. We use exactly the same clinical and laboratory information that the doctor currently uses to make a diagnosis.

We use a particular set of data to show how a medical diagnostic tool can be developed using the techniques of visual statistics. The data, which are described in detail below, were obtained from patients examined by a dermatologist about problems with their skin. The data include information obtained during an office visit and the results of a laboratory analysis of a biopsy taken from a patient.

When a doctor diagnoses a patient, he or she is assigning a disease category to the patient, using information obtained in the office and from the lab. If a group of possible diagnoses are all quite distinctly different in their symptomology, the diagnosis will be straightforward and will probably be correct. However, if a group of diseases share a number of similar features, the doctor's decision will be more difficult, may require knowledge about a wider array of information, and will be more likely to be incorrect. Of course, if the patient is diagnosed incorrectly , actually having a different condition than the condition diagnosed, the doctor's prescription may be inappropriate. Such misdiagnoses occur because the doctor must weigh very many sources of information and must decide on how much weight each source of information should get.

Frequently, doctors must choose between a diagnosis based on their personal clinical experience and a diagnosis based on laboratory tests. Of course, it is true that the doctor's subjective impressions and experiences are not totally reliable, but then neither are the laboratory tests. Although lab tests can be more accurate, they are not infallible, are often expensive, and can provide diagnoses whose validity is not self-evident. And it would be better if the doctor were able to weigh all of the evidence simultaneously, using both clinical and laboratory results together. But the amount of information that can be relevant can easily exceed even the smartest doctor's abilities, and the problem of how to weigh each of a large number of sources of information is beyond all of us.

**Table 2.1 Variables in the Dermatology Dataset**

| | |
|---|---|
| **Clinical Variables** | |
| *Erythema* | *Follicular papules* |
| *Scaling* | *Oral mucosa involvement* |
| *Definite borders* | *Knee and elbow involvement* |
| *Itching* | *Scalp involvement* |
| *Koebner phenomenon* | *Family history* |
| *Polygonal papules* | *Age* |
| **Histopathological Variables** | |
| *Melanin incontinence* | *Pongiform pustule* |
| *Eosinophils in the infiltrate* | *Munro microabcess* |
| *PNL infiltrate* | *Focal hypergranulosis* |
| *Fibrosis of the papillary dermis* | *Disappearance of the granular layer* |
| *Exocytosis* | *Vacuolization and damage basal layer* |
| *Acanthosis* | *Spongiosis* |
| *Hyperkeratosis* | *Saw-tooth appearance of retes* |
| *Parakeratosis* | *Follicular horn plug* |
| *Clubbing of the rete ridges* | *Perifollicular parakeratosis* |
| *Elongation of the rete ridges* | *Inflammatory mononuclear infiltrate* |
| *Thinning of Suprapapillary epidermis* | *Bandlike infiltrate* |

As seen by statisticians, medical diagnosis is one of many situations in which we classify our new, unclassified data by comparing it with benchmark datasets for which classifications exist, selecting the most similar benchmark and using its classification as the classification for our newly acquired data. Note that the benchmarks may be actual sets of empirical data that have somehow already been classified, or the benchmarks may be theoretical, prototypical datasets that are the idealized exemplars of the classification. It doesn't matter.

**Data**. The data are observations of 34 variables obtained from 366 dermatology patients. The variables are listed in Table 2.1. Twelve of the variables were measured during the office visit, 22 were measured in laboratory tests performed on a skin biopsy obtained during the office visit. Of the 34 variables, 32 were measured on a scale running from 0 to 3, with 0 indicating absence of the feature and 3 the largest amount of it. Of the remaining two variables, *Family history* is binary and *Age* is an integer specifying the age in years. Eight of the patients failed to provide their age. These patients have been removed from the analysis. The data were collected by Nilsel Ilter of the University of Ankara, Turkey (Guvenir et al., 1998).

Two difficulties must be dealt with before we can visualize these data: (1) The data are discrete—All of the variables except *Age* have four or fewer observation catego-

ries; and (2) There are too many variables—humans can not understand 34 variables simultaneously.

If we picture the data directly, we soon see the problems. For example, a matrix of scatterplots of the first three variables is shown in Figure 2.5. A scatterplot matrix has variable names on the diagonal and scatterplots off-diagonal. The scatterplots are formed from the variables named on the diagonal of a plot's row and column. The upper-left triangle of the matrix is a mirror image of the lower right. Scatterplot matrices are discussed in more detail in Chapter 7.

The discrete nature of the variables means that we have only a few points showing for each scatterplot, and that they are arranged in a lattice pattern that cannot be interpreted. For each plot, each visible point actually represents many observations, since the discrete data make the points overlap each other. In essence, the resolution of the data, which is four values per variable, is too low. The fact that there are 34 variables means that the complete version of Figure 2.5 would be a $34 \times 34$ matrix of scatterplots, clearly an impossibly large number of plots for people to visualize. Here, the problem is that the dimensionality of the data, which is 34, is way too high. Thus, the data cannot be visualized as they are, because their resolution is too low and their dimensionality is too high.

**Principal components.** All is not lost! These two problems can be solved by using principal components analysis (PCA). PCA reduces the dimensionality of data that have a large number of interrelated variables, while retaining as much of the data's original information as is possible. This is achieved by transforming to a new set of variables, the *principal components*, which are uncorrelated linear combina-

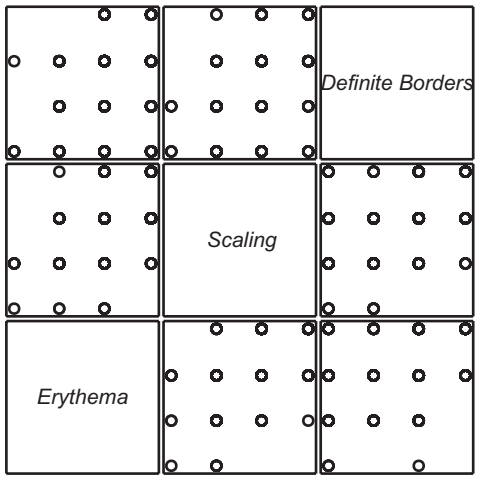



**Figure 2.5 Scatterplot matrix for three variables.**

tions of the variables. There is no other set of $r$ orthogonal linear combinations that fits more variation than is fit by the first $r$ principal components (Jollife, 2002).

Principal components have two important advantages for us. First, since only a few components account for most of the information in the original data, we only need to interpret displays based on a few components. Second, the components are continuous, even when the variables are discrete, so overlapping points are no longer a problem.

Figure 2.6 shows a scatterplot matrix of the five largest principal components. These components account for 63% of the variance in the original 34 variables. The general appearance of Figure 2.6 suggests that the observations can be grouped in a number of clusters. However, the total number of clusters as well as their interrelationships are not easily discerned in this figure because of the limited capabilities of

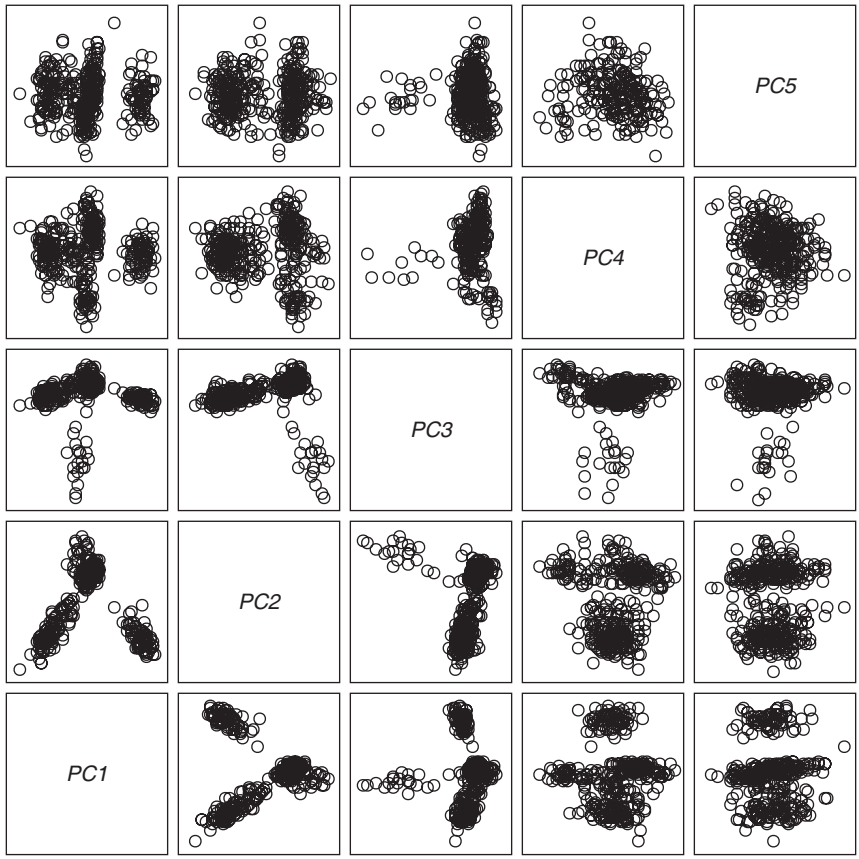



**Figure 2.6 Scatterplot matrix for the first five principal components.**

scatterplot matrices and because of the small plot sizes. There are ways around both of these limitations, as we discuss in Chapter 8.

**Linking.** Linking is a powerful dynamic interactive graphics technique that can help us better understand high-dimensional data. This technique works in the following way: When several plots are linked, *selecting* an observation's point in a plot will do more than highlight the observation in the plot we are interacting with—it will also highlight points in other plots with which it is linked, giving us a more complete idea of its value across all the variables. Selecting is done interactively with a pointing device. The point selected, and corresponding points in the other linked plots, are highlighted simultaneously. Thus, we can select a cluster of points in one plot and see if it corresponds to a cluster in any other plot, enabling us to investigate the high-dimensional shape and density of the cluster of points, and permitting us to investigate the structure of the disease space.

**Interpretation.** Figure 2.7 displays a "comic book"-style account of the process of selecting the groups of skin diseases that were visible in Figure 2.6. Frames of Figure 2.7 are the scatterplots of PC1 versus PC2 to PC5. The frames are to be examined sequentially from left to right and from top to bottom. The last frame is a repetition of the first frame and is intended to show the final result of the different actions carried out on the plots. An explanation of the frames follows.

1. PC1 vs. PC2: This scatterplot shows quite clearly three groups of points, labeled A, B, and C. Two groups are selected, but the third one remains unselected.
2. PC1 vs. PC3: Three actions are displayed in this frame. The groups selected in the preceding frame have been marked with symbols: A has received a diamond ($\Diamond$), and B a cross (+). Also, we can see that dimension PC3 separates group C into two parts: one compact, the other long. We have selected the observations in the long part and called them cluster $C_1$. At this point we have four clusters: A, B, $C_1$, and unnamed.
3. PC1 vs. PC4: This plot has the points selected in the preceding frame represented by a square symbol ($\square$). Unassigned observations above and below of squares make two groups. We selected the group with positive values in PC4 and called it $C_2$, giving us five clusters: A, B, $C_1$, $C_2$, and unnamed.
4. PC1 vs. PC5: We assigned the symbol ($\times$) to the group $C_2$ and selected the top values of PC5. Notice that this selection involved the reclassification of some points that had previously been assigned to the group $C_2$. The points selected will define group $C_3$. Notice that there are still some points that keep the original symbol of the points in the plot [a disk ($\bigcirc$)]. We call it group $C_4$. We now have six clusters: A, B, $C_1$, $C_2$, $C_3$ and $C_4$.
5. PC1 vs PC2 again: This frame is the same as the first frame in the sequence except that it shows the plot after steps 1 to 4. Note that we ran out of easily seen symbols for marking $C_3$, so we used gray to identify points. This frame

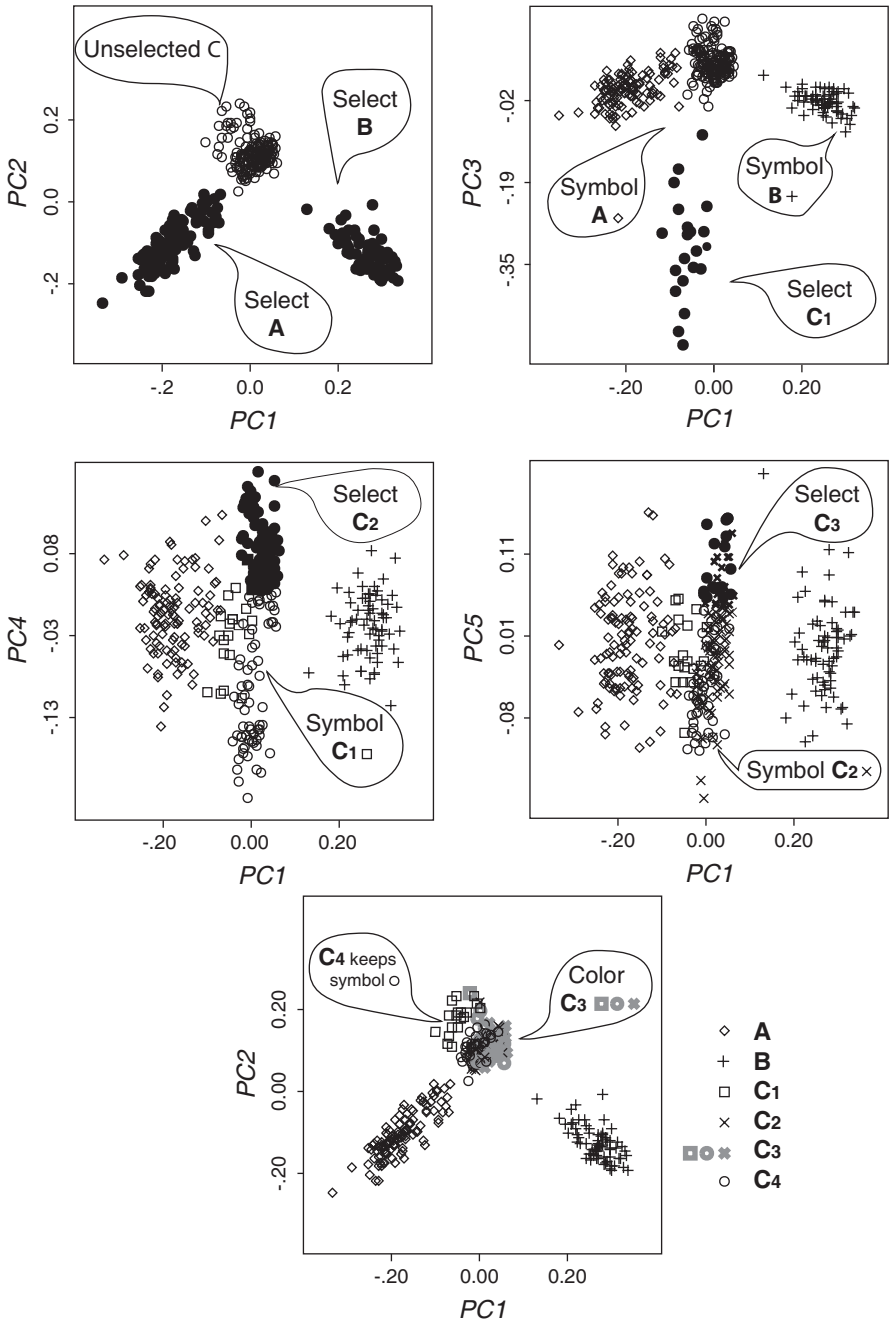



**Figure 2.7 Steps in selecting and changing symbols and colors in PCs.**

displays very clearly the three big clusters identified at the beginning and also suggests something of the finer structure inside cluster C.

A downside of the plots in Figure 2.7 is that the four clusters identified as $C_1$ to $C_4$ are not visualized very clearly. This problem suggests using *focusing,* a technique described in Chapter 4, to remove from the plot the points in the two largest clusters. As the remaining four subclusters are defined basically using PC3 to PC5, it makes sense to use a 3D plot to visualize them. Figure 2.8 displays such a plot after rotating it manually to find a projection clearly displaying the four clusters. This view suggests that some points actually seem to belong to clusters other than those to which they had previously been assigned. Thus, we reassign them, as explained in the balloons.

**Validity.** We did not mention it earlier, but the data include a diagnosis made by the doctor of each patient. It is interesting to compare our visual classification with the diagnostic classification. Table 2.2 presents a confusion matrix, which is a table that shows the frequency with which members of each of our classes were assigned to each of the diagnostic classes. If our visual classification agrees exactly with the diagnostic classification, the confusion matrix will be diagonal, there being zeros in all off-diagonal cells. To the extent that our visual classification is "confused," there will be nonzero frequencies off the diagonal.

In general, we see that our visual classes correspond closely to the diagnostic classes. All of the patients diagnosed with psoriasis and lichen planus were classified visually into groups A and B. The observations in the cluster labelled C are very well separated with respect to clusters A and B, with only one observation out of place. However, some of the subclusters in C, especially $C_2$ and $C_3$ have considerable interchanges between them, suggestion that additional efforts for improving the discrimi-

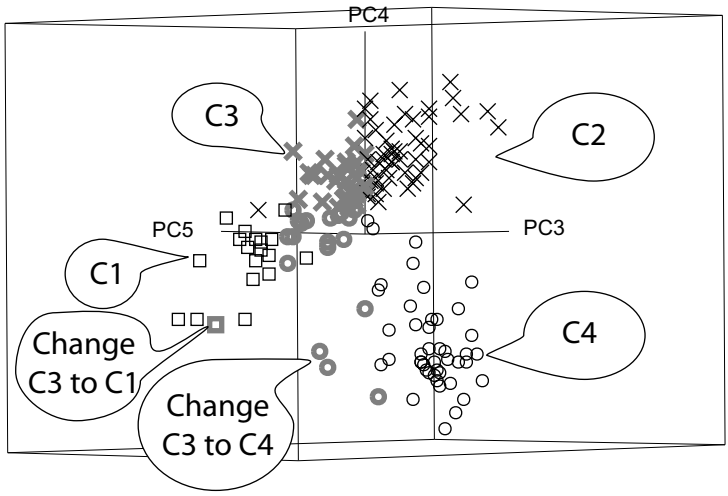



**Figure 2.8 Spinplot showing observations in clusters $C_1$ to $C_4$**

**Table 2.2 Confusion Matrix for the Medical Diagnosis Data**

| | A | B | $C_1$ | $C_2$ | $C_3$ | $C_4$ | S |
|---|---|---|---|---|---|---|---|
| Psoriasis | 111 | | | | | | 111 |
| Lichen planus | | 71 | | | | | 71 |
| Pityriasis rubra pilaris | | | 19 | | 1 | | 20 |
| Pityriasis rosae | | | | 39 | 8 | 1 | 48 |
| Seborrheic dermatitis | 1 | | | 10 | 48 | 1 | 60 |
| Chronic dermatitis | | | | | | 48 | 48 |
| | 112 | 71 | 19 | 49 | 57 | 50 | 358 |

nation between them are necessary. This problem can result from the difficulties we had to correctly pinpoint the points in frame 4 of Figure 2.7. Of course, there could also be misdiagnoses by the doctor—we can't tell. Nevertheless, 93.8% of the cases are classified correctly which is only 2.4% lower than the classification made originally using a special algorithm called VFI (Guvenir et al., 1998).

## 2.3 Fidelity and Marriage

Fitting statistical models involves finding explanations or simple descriptions for patterns in data. Typically, there is an outcome or response variable, and one or more predictors or explanatory variables that can be used individually or jointly (as in an interaction) for this purpose. The goal usually is to find the simplest adequate description—the smallest, least complex model—which nonetheless provides a reasonably complete summary of the data. Any statistical model can be cast as a breakdown of the data into two parts: what we have summarized, described, or explained (called the *model*) and the rest, which we do not understand (the residual). That is,

$$data = model + residual$$

You can always make the model fit perfectly (by including as many parameters as you have data values), or make the model incredibly simple (just choose an empty model), but the payoff comes when you can find a balance between fit and parsimony.

Traditional statistical methods use various numerical measures of goodness of fit [$R^2, \chi^2$ and parsimony (degrees of freedom)], and often combine these to determine optimal points along a trade-off relation (adjusted $R^2$, AIC, etc.). We prefer to combine various sources of model information into coherent displays, fitting models visually by direct manipulation. We call this paradigm *visual fitting*.

We illustrate this approach with an example of fitting loglinear models to data about divorce and the occurrence of premarital or extramarital sex. The data, a cross-classified table of frequencies, are shown in Table 2.3. Thornes and Collard (1979) obtained two samples of roughly 500 people, one of married people, the other of those seeking divorce. Each person was asked (a) whether they had sex before marriage, and (b) whether they had extramarital sex during marriage. When these are broken down by gender, we get the $2 \times 2 \times 2 \times 2$ frequency table. A log-linear model

**Table 2.3 Cross-tabulation of Marital Status Versus Premarital Sex, Extramarital Sex, and Gender**

| *Premarital* | *Extramarital* | *Gender* | *Marital Status* | |
|---|---|---|---|---|
| | | | Married | Divorced |
| Yes | Yes | Male | 11 | 28 |
| | | Female | 4 | 17 |
| | No | Male | 42 | 60 |
| | | Female | 25 | 54 |
| No | Yes | Male | 4 | 17 |
| | | Female | 4 | 36 |
| | No | Male | 130 | 68 |
| | | Female | 322 | 214 |

attempts to explain the pattern of frequencies in this table, so-named because it uses a linear model of the logarithm of the frequency.

The question we wish to ask of these data is: How is *Marital Status* related to (how does it depend on) *Gender*, *Premarital Sex* and *Extramarital Sex*? Before we go any further, ask yourself these questions. What factors influence the likelihood of divorce? Do any have combined (interactive) effects?

We can translate this question into a form that can be used for modeling by a log-linear statistical model as follows. In terms of understanding *Marital Status*, the explanatory variables—*Gender*, *Premarital Sex,* and *Extramarital Sex*—can be associated in any arbitrary ways; they just describe the sample. Probably, men are more likely to have had premarital sex than women—a *(GP)* association—and maybe also more likely to have had extramarital sex *(GE)*. We don't really care, and lump all associations among *G*, *P*, and *E* into one term, *(GPE)*.

**Empty model.** The basic, empty model is symbolized as *(GPE)(M)*, and asserts that *Marital Status* has no association with *Gender*, or with *Pre-* or *Extramarital Sex*.

**Saturated Model.** At the opposite end of the parsimony spectrum is the saturated model, (*GPEM*). This model allows marital status to be associated with all of the factors and their combinations, in unknown ways. However, it always fits perfectly, so it cannot possibly tell us anything.

**Model-Fitting strategies.** What we must do is find a model somewhere in between these two extremes, a model that is parsimonious, yet explanatory. That means that we must search through a variety of models, trying to decide which is best. There are two basic strategies for doing this search:

- **Forward search**. start with the empty model (which will fit poorly), and add terms allowing associations of *M* with *G*, *P*, and *E* and their combinations, until the model fits well enough.

- **Backward Search**. Start with the saturated model and remove associations of *M* with the others, as long as the model does not fit too poorly.

We use the former method here to illustrate visual fitting. The spreadplot (a kind of multiplot visualization that is introduced in chapter 4) for the initial model, *(GPE)(M)* is shown in Figure 2.9 (on the following two pages). This model fits very poorly, of course ($G^2$=107, df = 7, p < 0.001). The $G^2$ measure is a badness-of-fit measure. Low values are good, high values are bad. The empty model, reported here, has a very large value of $G^2$, meaning the fit is very poor, which, of course, it must be, since it has no terms. The hypothesis test, when rejected, as is the case here, indicates the model does not fit the data.

However, the spreadplot is not limited to displaying goodness-of-fit statistics. It also reports in a graphical way other aspects of importance in the model fitting situation. These other elements are indicated in the spreadplot by means of text inside balloons.

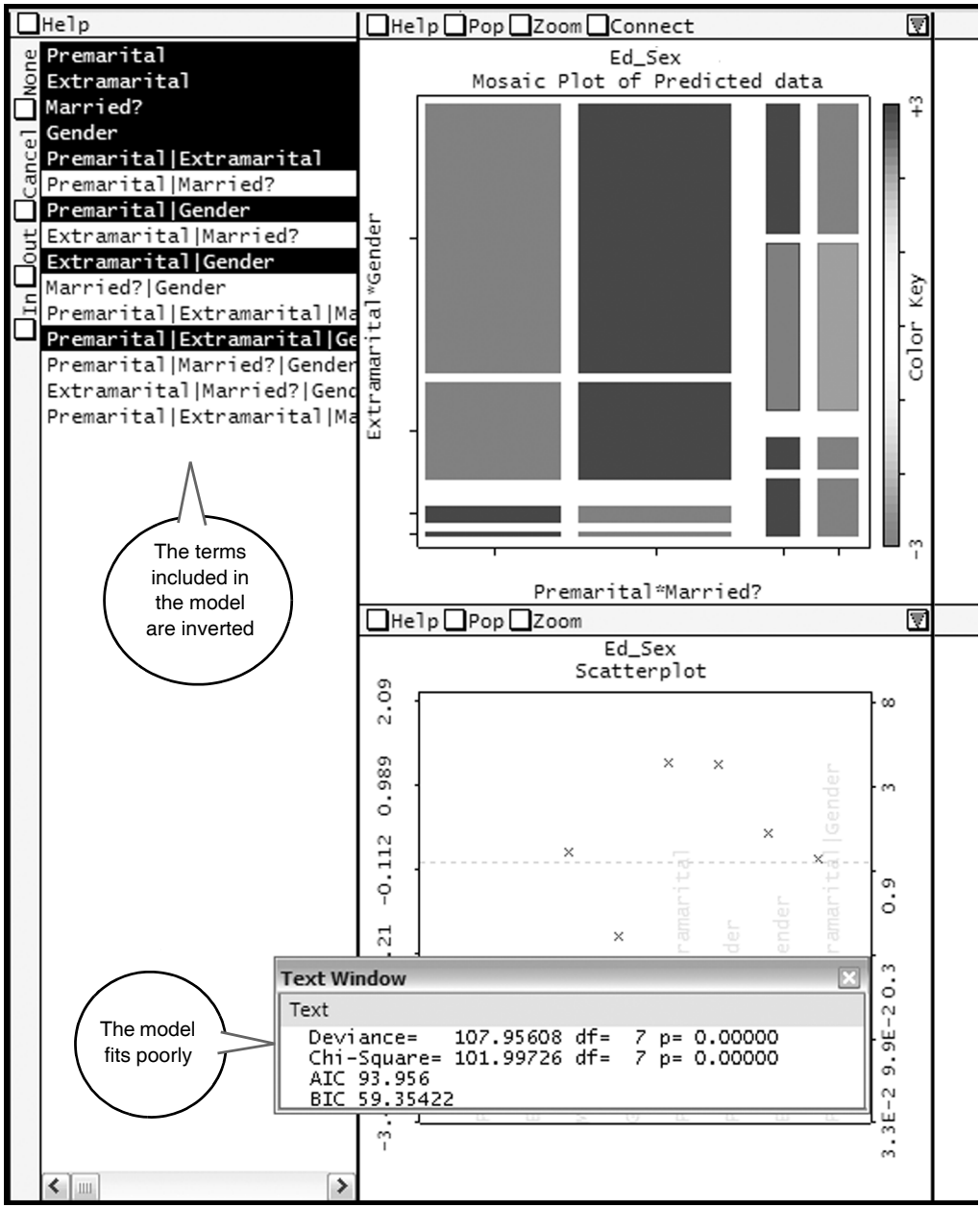
The way that the model differs from the data gives us clues about how we can improve our model. We can use mosaic displays to find the specific ways in which the model is different from the data, since mosaics show the residuals (or differences) of the cells with respect to the model. Looking at these differences, we can observe patterns in the deviation that will help us in our search.

Unfortunately, mosaic displays are best viewed in color, and we are forced to use black and white. (We do the best we can, but to be honest, the black-and-white versions shown in Figure 2.9 do not do justice to the mosaic displays. If you can view this online, please do; it will help). So while the color versions of mosaic displays use two different colors for representing positive or negative residuals (and the hue of the color indicates their absolute values), the black-and-white version puts everything in gray. Therefore, we have made special versions of the mosaic displays to use for this book.

Figure 2.10 displays black-and-white mosaic displays for the initial model *(GPE)(M)*. The displays represent positive residuals in black and negative residuals in white. Although this way of coding has the disadvantage of ignoring the actual value of the residuals, we believe that this disadvantage is not very consequential, as, in practice, the analyst pays attention to the patterns of signs of the residuals, such as shown in these black-and-white versions of mosaic displays. Of course, problems can arise with small residuals because they need to be assigned to one color or another. As a solution, we specify cutoff values that we find appropriate for each situation. Values above the absolute values of the cutoff will be displayed in black or white (depending on their sign), and the rest in gray. We used 3 as the cutoff value in Figure 2.10 and as the model fits quite poorly, all the values were above it in absolute values. Thus there is no gray.

There are three mosaic displays in Figure 2.10. Figure 2.10a is the initial display produced by the software. The order of the variables in this display is *PEMG* and was defined simply by the original order of the variables in the datafile. Even though this display looks quite simple and we could possibly interpret it, exploration of the spe-

cific values of the residuals suggested that we change the order of the variables in the display to make it more interpretable.



Let's look at the residuals table below Figure 2.10a. The first column of this table is the negative of the second, and similarly for the third and fourth. This is a consequence of the marginal constraints for this model, which make some residuals redundant (Agresti, 1990). Comparing the first and second columns or the third and fourth,

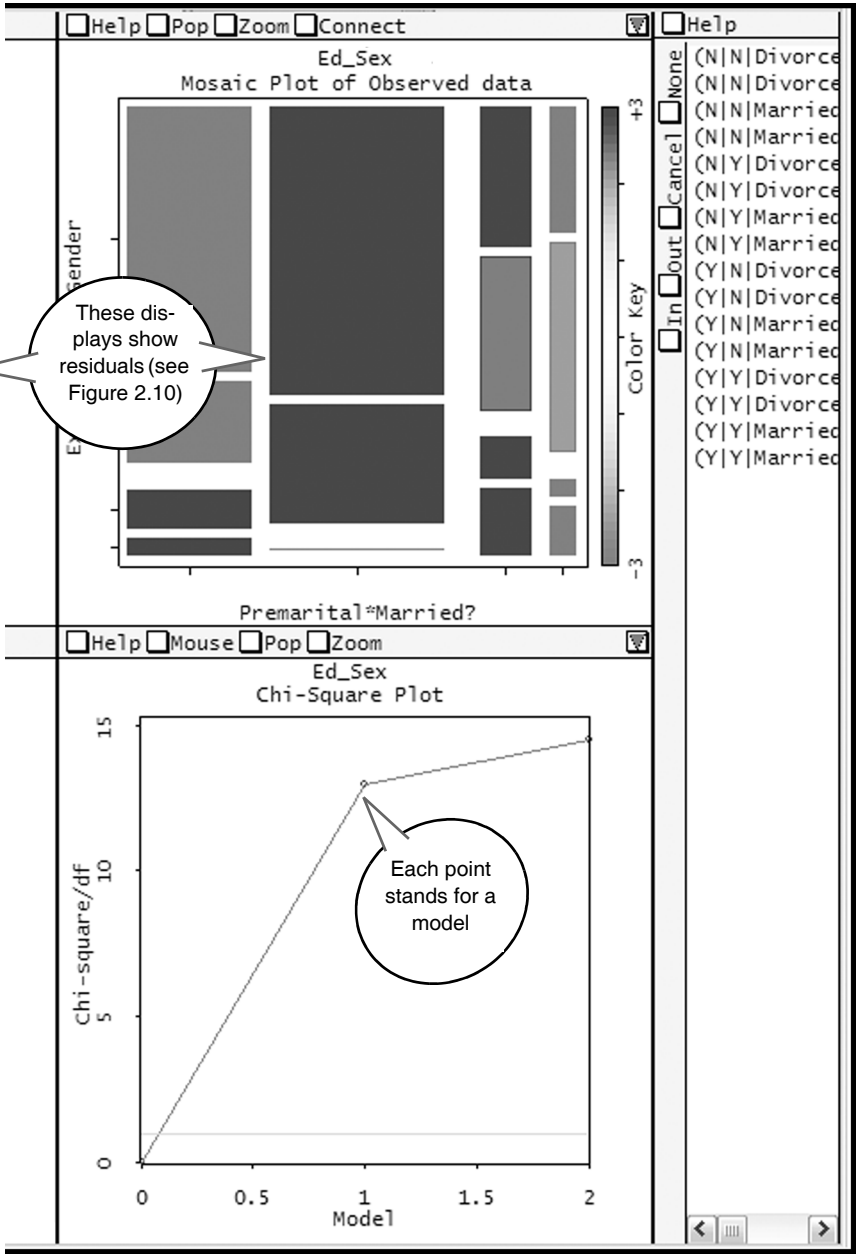



**Figure 2.9 Spreadplot for the initial model—the empty model (*GPE*)(*M*).**

is therefore uninformative. More interesting conclusions may be extracted by comparing the first and third columns in Figure 2.10a (or the second and the fourth).

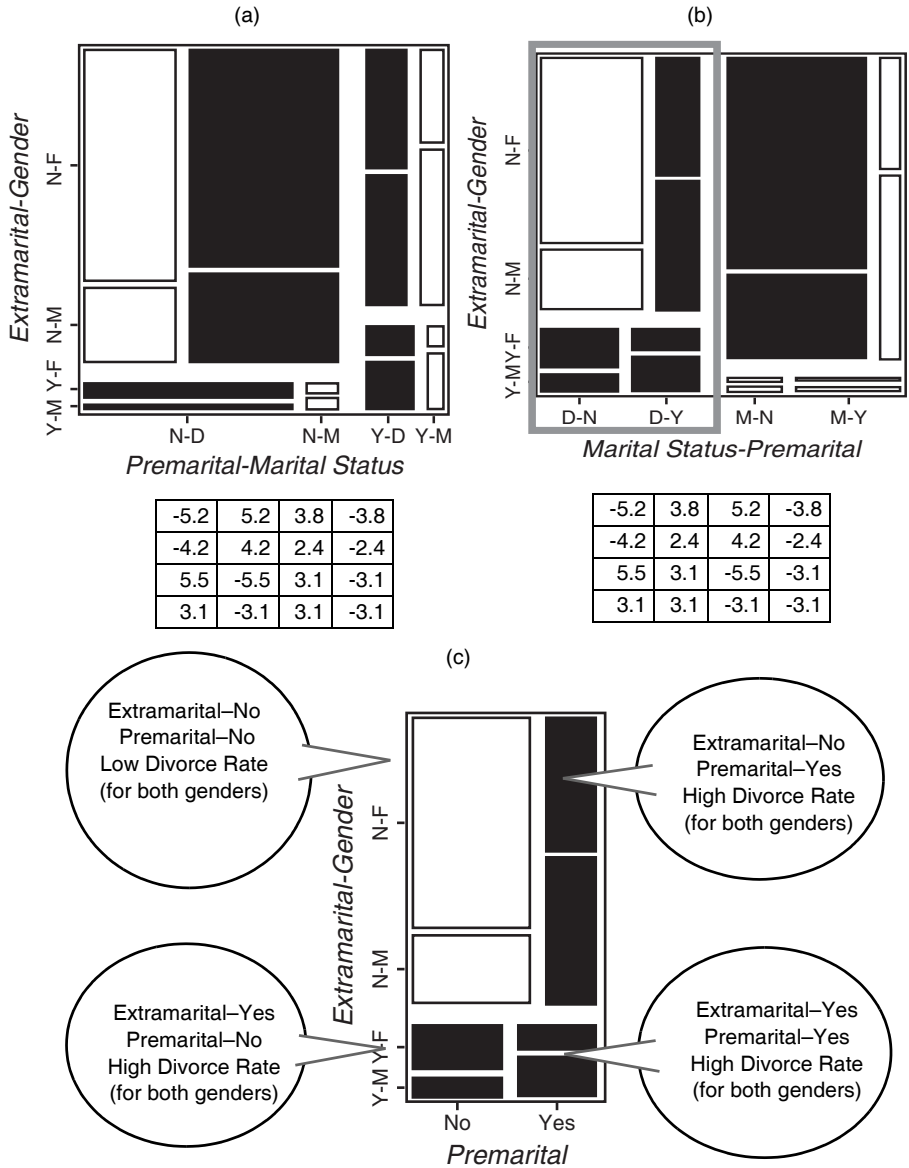



**Figure 2.10 Residual patterns for the model *(GPE)(M)*: (a) Mosaic display with the default order of the variables in the software *PEMG;* (b) using order *MEPG*; (c) focused view of rectangular section of (b) (*Marital Status* = Divorced).**

Manipulating the mosaic display in Figure 2.10a we transformed it to the mosaic display in Figure 2.10b. Notice that the first and third columns in Figure 2.10a are now the first and the second, and that second and fourth are the third and fourth. This was attained by exchanging the order of the variables *PreMarital* and *Marital Status* so that the order of the variables in the display is *MEPG*. Now we can see that the two initial columns mirror the other two columns (i.e., the divorced part of the display is a mirror of the married part). Hence, we can focus (*focusing* is a technique discussed in Chapter 4) on these two columns without missing any information, as the other part is loaded with exactly the same information. The part we have selected is marked with a rectangle in Figure 2.10b and is displayed separately in Figure 2.10c with balloons that describe the results. As can be seen in this figure, the information can be summarized in only a sentence: people with sexual encounters (*Pre-* or *Extra-*) out of marriage are more likely to divorce than those without them. This statement suggests the model *(GPE)(PM)(EM)*, which asserts a relationship between *Marital Status* and *Pre-* and *Extramarital Sex*, to be tested next.

Figure 2.11a shows the two first columns of the mosaic display for the model *(GPE)(PM)(EM)* with the variables in the same order as in Figure 2.10c. This model still does not fit ($G^2 = 18.15$, df = 5, $p \approx 0.002$) but improves the basic model considerably. In this mosaic display, residuals in absolute value larger than 1 were filled in black or white (depending on the sign of the residual) and the rest were filled in gray. Notice that there are only two black (positive) residuals in the display as well as only two white (negative) residuals. Again, there is evidence of some regularity in the display but as we did previously, we chose to manipulate the mosaic to make it easier to interpret: Thus, we exchanged the order of *Extramarital Sex* and *Gender* variables in

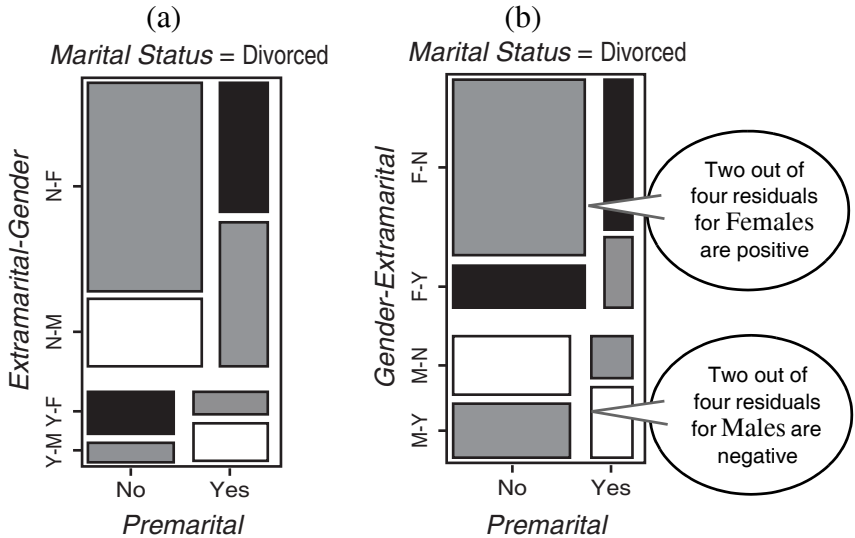



**Figure 2.11 Mosaic displays for model *(GPE)(ME)(MP)* focused on *Marital Status* = Divorced. (a) Mosaic display using order *MEPG*; (b) using order *MGPE*.**

the plot. The result is shown in Figure 2.11b. The visual effect of this manipulation is that the two positive residuals are located in the upper part of the plot, and the negative residuals are in lower part of the plot. Looking at the labels, we can see that the positive residuals are related with females, and negative residuals are related with males (i.e. females are more prone to divorce than males). This points to a relation between the variables *Gender* and *Marital Status* and suggests the model *(GPE)(PM)(PE)(GM)* as the next one to be tested.

The improvement of model *(GPE)(PM)(PE)(GM)* with respect to the model without the *(GM)* term was modest ($G^2 = 13.62$, df=4, $p \approx 0.008$). This suggests that we compare models to see if the difference is important. We can do this comparison using the Chi-Square plot shown in the spreadplot of Figure 2.9 by selecting the points corresponding to the two models that we want to compare. The result ($\Delta G^2 = 4.53$, df=1, $p \approx 0.03$) suggests that the difference is larger than zero, but not by much. Actually, we used our software to put the mosaic displays for both models side by side as well. These are shown as Figure 2.12a and b. In these displays, residuals larger than 1 are shown in black, and residuals smaller than 1 in white (the rest are in gray). A third display, Figure 2.12c, shows the *differences* between the absolute value of residuals of the other two models and can be used to explore which cells of the second model have actually improved the fit out of the first model. So cells in black in this display are those that have larger residuals in the second model than in the first, and cells in

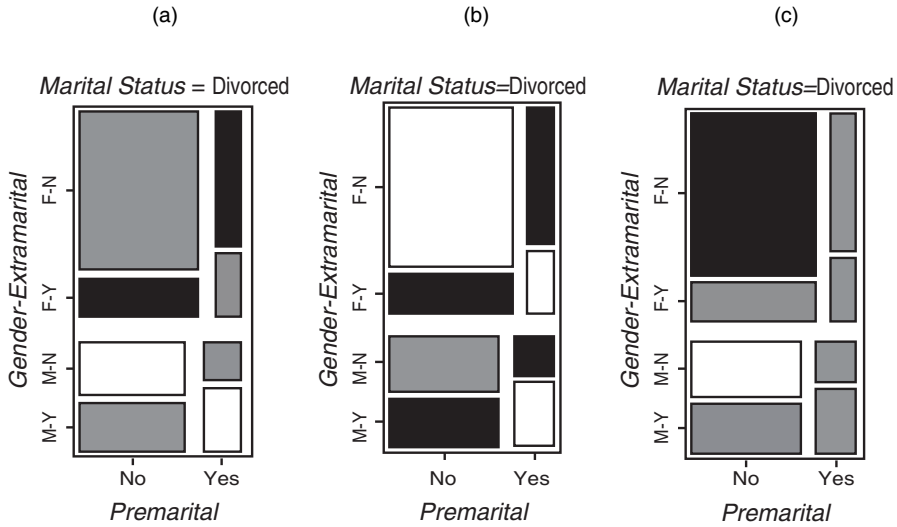



**Figure 2.12 Mosaic displays before and after adding the term *(GM)* to the model *(GPE)(ME)(MP)* focused on *Marital Status* = Divorced. (a) Mosaic display for model *(GPE)(ME)(MP)*; (b) mosaic display for model *(GPE)(PM)(PE)(GM*; (c) mosaic display of differences between (a) and (b).**

white are those that have smaller residuals (in both cases, the difference must be larger than 1 to use this coding). Cells in gray are for differences not larger than 1.

The visual impression of mosaic displays in Figure 2.12 is that adding the *(GM)* term does not improve the fit at all. On the contrary, the fit of the model seems to worsen, as seven out of eight cells have residuals larger than 1. Looking at the mosaic display of differences in Figure 2.12c we can see that actually only two cells have changed its fit more than 1: males without *Pre-* or *Extramarital* encounters (which have reduced the residual) and females (which have increased the residual). In conclusion, the apparent interaction between *Gender* and *Marital Status* that we observed in Figure 2.11 seems irrelevant at this stage of the analysis of our data.

We will look at Figure 2.12a again to find hints for new terms in the model. The current layout of this plot hinted that females had more divorces than males, but it is possible that a change in the aspect of this display will bring about other suggestions. Figure 2.12a suggests some type of regularities that we can not easily see at this point. Manipulating the order of the variables we arrived at Figure 2.13a, which has the right layout to see a new pattern in the residuals. In this figure, the variables *Pre-* and *Extramarital* are set together on one axis of the plot, and the variables *Marital State* and *Gender* are on the other. In this layout, the two negative residuals (white cells) are in the sides of the plot, and the two positive residuals (black cells) are in the center. If we combine the residuals by *Gender,* as shown by the dashed rectangle in Figure

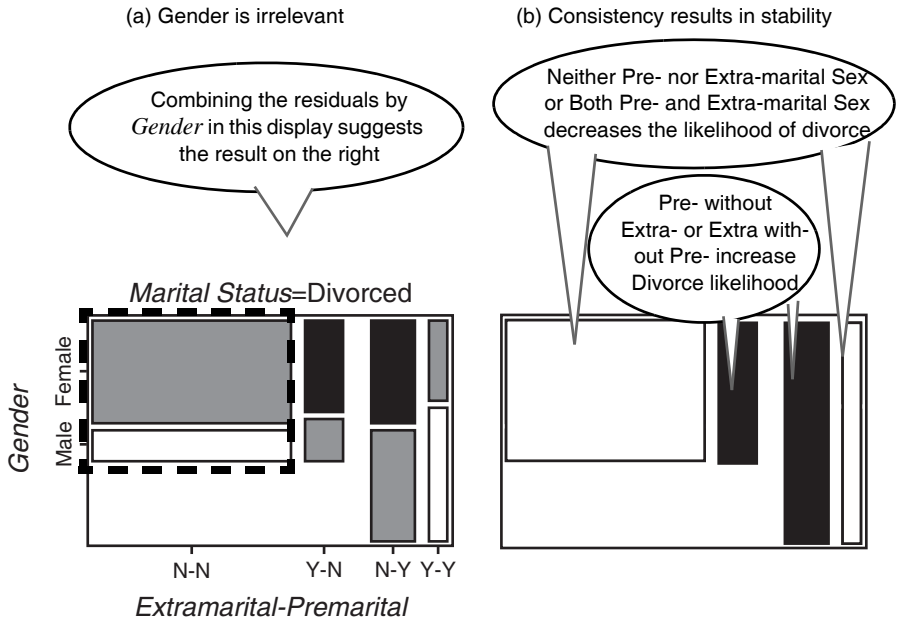



**Figure 2.13 Mosaic displays for model *(GPE)(PM)(PE)(GM)*. (a) Using order *PMEG* focused on *Marital Status*= Divorced; (b) Schema of display a when the residuals are combined by *Gender*.**

2.13a with those with no *Pre-* and no *Extramarital,* for the four combinations we can visualize the scheme in Figure 2.13b. This pattern suggests a three-way interaction among *Pre-* and *Extramarital* and *Marital Status (PEM).*

The model *(GPE)(PEM),* whic includes the interaction mentioned previously, fits the data well ($G^2 = 5.24$, df = 4, $p \approx 0.26$) and would be accepted as a good model when considering the usual goodness of fit criteria. Yet we can observe in the mosaic display for this model (Figure 2.14) a pattern that suggests that we can improve this model still more. Using a cut-off value of 0.8, the display shows that the females have higher likelihood of divorce than males given this model (constraints on this model make males and females have the same residuals in absolute value). We saw this trend previously, but we rejected it as unimportant because the effect of including this interaction in the model seemed too little. However, this display suggests testing the interaction in combination with the model currently defined.

The model *(GPE)(PEM)(GM)* fits very well ($G^2 = 0.69$, df = 3, $p \approx 0.87$) and even though this model is not very different from the model without the *(GM)* term ($\Delta G^2 = 4.53$, df = 1, $p \approx 0.03$), we regard this model as the most appropriate for our data.

The actions described in the previous paragraphs are very typical of the actions taken during a real data analysis. Normally, the nitty-gritty detail would be left out, but we think it is important to put them in so you can really learn about how to do this type of analysis. The typical cycle of analysis is this: The analyst evaluates the evidence to select the term that best explains the dependent variable. Then the analyst tests that term. If the result is an improvement, the variable or term is left in the equation, but if is not, the analyst may drop it and look at the displays for a new hint.

Thus, interactive dynamic graphics allows you to search for a good model using a process that is not strictly forward or backward, but involves both. The normal noninteractive stepwise approach is an automatic procedure that uses a statistical rule of thumb to add or delete predictors from the model. Often, automated stepwise proce-

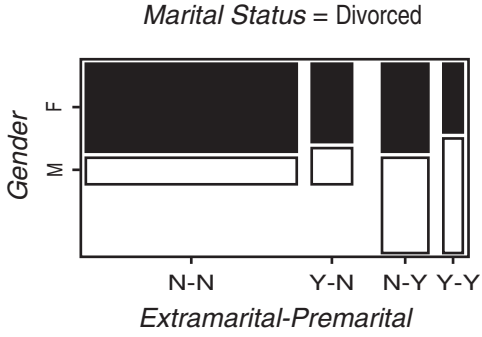



**Figure 2.14 Mosaic displays for model *(GPE)(PEM)* focused on *Marital Status*= Divorced. The order of the variables is *PMEG*.**

dures produce poor results because they are limited to simple rules and do not consider all the elements that an analyst would use. Consequently, many statisticians usually prefer to perform the modeling process by hand instead of using the automatic method. Visual fitting has the advantage of providing a concrete, perceptual goal: A good-fitting model will have small residuals and the mosaic will be mostly unshaded. Think of the search for an adequate model as "cleaning the mosaic."

However, the manual process also has a big disadvantage: It makes it difficult to record and describe the actual steps performed by the analyst. Indeed, it often happens that scientific papers or books only report the end model, without discussing the choices preceding the conclusion that such a model was the best. Software does not often help with this endeavor either, as it usually does not provide a way to overlook the models considered and rejected along the process. Fortunately, as we have already discussed in the spreadplot shown in Figure 2.9, we have a display in our software that addresses this problem.

Figure 2.15 shows a display of the fit of the models evaluated in this section. The display shows the $\chi^2/df$ value of each of the models considered. The horizontal line in the display stands for the rule of thumb that $\chi^2/df$ values below 1 can be considered models that fit well. The display has labels for each of the models we considered. As you can see, this display is an effective way to record and recall the entire set of models that we explored during a modeling session, such as the one we described in this section. And perhaps best of all, dynamic interactive graphics means that all we need to do to return to an earlier model is to click on its representation in the diagram. Then we are back to that model and can proceed from there as we wish.

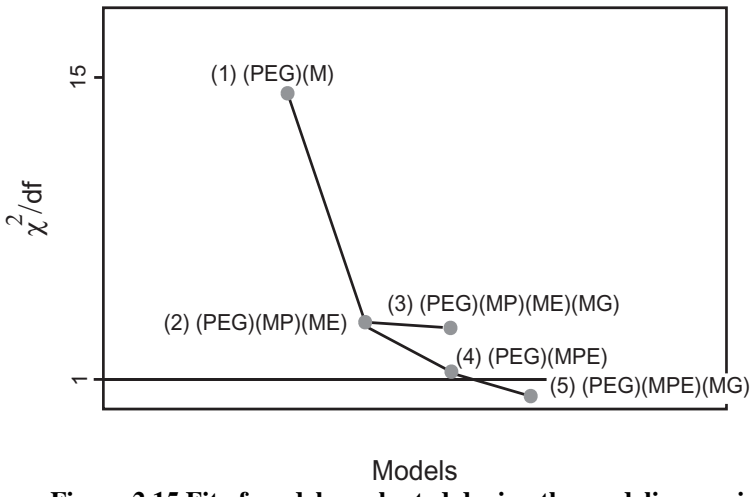



**Figure 2.15 Fit of models evaluated during the modeling session.**