Interface Foundation of America

Exploratory Data Analysis for Complex Models [with Discussion, Rejoinder]
Author(s): Andrew Gelman and Andreas Buja
Source: *Journal of Computational and Graphical Statistics*, Vol. 13, No. 4 (Dec., 2004), pp. 755-787
Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of America
Stable URL: http://www.jstor.org/stable/27594077
Accessed: 17-10-2017 05:21 UTC

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at http://about.jstor.org/terms

Institute of Mathematical Statistics, American Statistical Association, Interface Foundation of America, Taylor & Francis, Ltd. are collaborating with JSTOR to digitize, preserve and extend access to *Journal of Computational and Graphical Statistics*

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

Discussion Article

Exploratory Data Analysis for Complex Models

Andrew GELMAN

"Exploratory" and "confirmatory" data analysis can both be viewed as methods for comparing observed data to what would be obtained under an implicit or explicit statistical model. For example, many of Tukey's methods can be interpreted as checks against hypothetical linear models and Poisson distributions. In more complex situations, Bayesian methods can be useful for constructing reference distributions for various plots that are useful in exploratory data analysis. This article proposes an approach to unify exploratory data analysis with more formal statistical methods based on probability models. These ideas are developed in the context of examples from fields including psychology, medicine, and social science.

Key Words: Bayesian inference; Bootstrap; Graphs; Multiple imputation; Posterior predictive checks.

1. INTRODUCTION

This article proposes a unified approach to exploratory and confirmatory data analysis, based on considering graphical data displays as comparisons to a reference distribution. The comparison can be explicit, as when data are compared to sets of fake data simulated from the model, or implicit, as when patterns in a two-way plot are compared to an assumed model of independence. Confirmatory analysis has the same structure, but the comparisons are numerical rather than visual.

From the standpoint of exploratory data analysis, our methodology has three major benefits:

1. Explicit identification of a comparison model allows one to simulate replicated data to be used as a reference distribution for an exploratory plot.

Andrew Gelman is Professor, Department of Statistics and Department of Political Science, Columbia University, New York, NY 10027 (E-mail: gelman@stat.columbia.edu).

©2004 American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America
Journal of Computational and Graphical Statistics, Volume 13, Number 4, Pages 755–779
DOI: 10.1198/106186004X11435

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

2. Symmetries in the underlying model can be used to construct exploratory graphs that are easier to interpret, sometimes (as with a residual plot) without the need for explicit comparison to a reference distribution.
3. Inclusion of imputed missing and latent data can allow more understandable completed-data exploratory plots.

From the standpoint of complex modeling, our theory has the key benefit of suggesting exploratory plots that address the fit of data to whatever model is being fit, in comparison to the usual graphical methods that are constructed in the absence of an explicit model. In addition, the placement of exploratory data analysis within the general theory of model checking allows the potential for graphical methods to become a more automatic presence in statistical modeling.

Models have been evaluated by comparing real to simulated data for a long time (e.g., Bush and Mosteller 1955; Ripley 1988), and methods have also been developed for graphically assessing the fit of specific models (e.g., Landwehr, Pregibon, and Shoemaker 1984). This article attempts to encompass statistical graphics within the general theory of statistical inference to point out ways in which new graphical displays can be routinely developed and interpreted.

1.1 Background

In the past few decades, the scope of statistics has been broadened to include exploratory analysis and data visualization—going beyond the standard paradigms of estimation and testing, to look for patterns in data beyond the expected (see Tukey 1972, 1977; Chambers, Cleveland, Kleiner, and Tukey 1983; Cleveland 1985, 1993; Tufte 1983, 1990; Buja, Cook, and Swayne 1996; and Wainer 1997 among others).

At the same time, methods have been developed to fit increasingly complex and realistic models to data. The complex modeling methods include nonparametric and semiparametric methods, sieve models, tree-based models, and adaptive nonlinear regression approaches (for a review, see Hastie, Tibshirani, and Friedman 2002). This article focuses on parametric and Bayesian methods, where hierarchical models allow the fitting of high-dimensional models capturing heterogeneity, interactions, and nonlinearity; see, for example, Gelman, Carlin, Stern, and Rubin (1995), Carlin and Louis (1996), and Denison, Holmes, Mallick, and Smith (2002) for recent reviews.

Improvements in computation have spurred developments both in exploratory data analysis and in complex modeling. For exploratory data analysis and data visualization, higher-resolution graphics, more sophisticated interactive user interfaces, and more accessible software have given room for graphical methods to become more elaborate and also more widely available. For modeling, new algorithms ranging from neural networks to genetic algorithms to Markov chain simulation allow users to fit models that have no closed-form expressions for estimates, uncertainties, and posterior distributions. And, of course, both graphics and modeling have benefited from the sheer increase in the speed and storage capacity of desktop computers. The connections between statistics, graphics, and

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

computation appear even in the title of this journal.

Unfortunately, there has not been much connection made between research in the two areas of exploratory data analysis and complex modeling. On one hand, exploratory analysis is often considered in the absence of models. From the other direction, in Bayesian inference, exploratory data analysis is typically used only in the early stages of model formulation but seems to have no place once a model has actually been fit.

This article argues that (a) exploratory and graphical methods can be especially effective when used in conjunction with models, and (b) model-based inference can be especially effective when checked graphically. Our key step is to formulate (essentially) all graphical displays as model checks, so that new models and new graphical methods go hand-in-hand.

1.2 The Complementary Nature of Exploratory Data Analysis and Modeling

Exploratory analysis is often presented as model-free. However, the early article on the topic by Tukey (1972) focused on "graphs intended to let us see what may be happening over and above what we have already described," which suggests that these graphs can be built upon existing models. Tukey contrasted exploratory analysis with calculations of p values, or *confirmatory data analysis*. These two sets of methods are both forms of model checking: exploratory data analysis is the search for unanticipated areas of model misfit, and confirmatory data analysis quantifies the extent to which these discrepancies could be expected to occur by chance. The exploratory methods of Tukey tend to be based on fairly simple models such as additive fits and the Poisson distribution for counts; we would like to apply the same principles to the more complex models that can be fit today using methods such as Bayesian inference and nonparametric statistics.

In some standard model diagnostics, the connection between exploratory and confirmatory analysis is clear: for example, in a quantile-quantile plot (Wilk and Gnanadesikan 1968), the shape of the line shows the discrepancy between the empirical distributions of model and data (or between two datasets), and the magnitude of the discrepancies from the 45° line can be calibrated at any given level of statistical significance based on simulation or the theory of order statistics. More recently, Buja and Rolke (2003) showed how to calibrate such tests to account for simultaneous comparisons.

More generally, complex modeling makes exploratory analysis more effective in the sense of being able to capture more subtle patterns in data. Conversely, when more complex models are being used, graphical checks are more necessary than ever to detect areas of model misfit.

- On a practical level, we suggest to modelers that they check the fit of their models using simulation-based model checks—comparisons of observed data to replications under the model. Conversely, we suggest to exploratory data analysts that they proceed iteratively, fitting as much structure as possible into a model and then using graphs to find patterns that represent deviations from the current state-of-the-art model. In addition, we suggest

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

applying the methods and ideas of exploratory data analysis to structures other than raw data, such as plots of parameter inferences, latent data, and completed data.

- On a theoretical level, we identify different sorts of graphical displays with different symmetries or invariances in an explicit or implicit reference distribution of test variables. In one direction, this is an attempt to put some theoretical structure on graphics and exploratory data analysis, so that existing and proposed graphical methods can be interpreted in terms of implicit models. In the other direction, this theory is intended to give some guidelines into how to most effectively express a model check as a graphical procedure.

Section 2 of this article reviews the Bayesian theory of posterior predictive model checking (which we recommend as a method for interpreting statistical graphics even if Bayesian inference is not being used to fit the model). Section 3 applies this theory to categorize exploratory graphs in a modeling context. Section 4 concludes with a discussion of the steps that would be required to integrate graphics with complex modeling in routine statistical practice.

2. STATISTICAL GRAPHICS AS MODEL CHECKING

We view *model checking* as the comparison of data to replicated data under the model. This includes both exploratory graphics and confirmatory calculations. In either case, our goal is not the classical goal of identifying *whether* the model fits or not (and certainly not the goal of classifying models into correct or incorrect, which is the focus of the Neyman-Pearson theory of Type 1 and Type 2 errors), but rather to understand in what ways the data depart from the fitted model.

From this perspective, the two key components to an exploratory model check are (1) the graphical display and (2) the reference distribution to which the data are compared. As we discuss in Section 3, the appropriate display depends on the aspects of the model being checked. This section discusses the reference distribution, that is, the procedure for generating replicated datasets y^{rep} to be compared to observed data y.

2.1 Posterior Predictive Checking

Consider data y that are used to estimate a parameter vector $\boldsymbol{\theta}$. For the purpose of model checking, the data are summarized by a test statistic $T(y)$ (which can be a graphical display, as discussed in Section 3.2). The posterior predictive distribution, $p(y^{\text{rep}}|y)$, represents the distribution of future data under the model being fit. Checks of model fit can be framed as comparisons of $T(y)$ to the replication distribution of $T(y^{\text{rep}})$ under the posterior predictive distribution. From the Bayesian standpoint, the parameters, data, and replicated data have a joint distribution, $p(y, y^{\text{rep}}, \boldsymbol{\theta})$ which is symmetric in y and y^{rep} (Meng 1994; Gelman 2003).

The statistical literature features a variety of ways of defining reference distributions,

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

including permutation tests, bootstraps (Efron and Tibshirani 1993), cross-validation (Stone 1974; Gelfand, Dey, and Chang 1992), and posterior predictive checks (Rubin 1984; Gelman, Meng, and Stern 1996). Buja, Asimov, Hurley, and McDonald (1988, sec. 5) considered several methods of constructing reference distributions for visual comparisons, including the bootstrap and the parametric bootstrap. For example, if $\boldsymbol{\theta}$ is estimated by maximum likelihood, it might be convenient to sample y^{rep} from the distribution $p(y|\hat{\boldsymbol{\theta}})$, which we would view as an approximate posterior predictive distribution. Section 3 also considers implicit reference distributions and symmetry properties of reference distributions that are not fully specified (e.g., regression residuals that are modeled as independent with zero mean but with no necessary distributional specification).

In general, the reference distribution corresponds to a model for data generation. For the purposes of this article, it is not so important where the reference distribution comes from, just that it has been defined in some way. This general perspective was implicit in Tukey (1977) and was more explicitly presented by Finch (1979) and Buja et al. (1988), who considered the application of permutation tests even in the absence of probability models or randomization. As is common in statistics, various standard approaches can be derived in more than one possible way. For example, the binomial distribution, which has long been a reference distribution for count data (see, e.g., Stigler 1986), can also be derived as a conditional form of the Poisson distribution or as a permutation distribution from a superpopulation of trials.

In the standard theory of posterior predictive model checking (e.g., Gelman, Meng, and Stern 1996), the test statistic $T(y)$ is a scalar, and its discrepancy with the model is summarized by a p value, $\Pr(T(y) > T(y^{\text{rep}})|y)$. Here, however, we are interested in vector test statistics $T(y)$, which can be displayed graphically; that is, the graph itself is the test statistic, and its replication distribution is indicated by several random simulations of this graph, as it might appear in replicated datasets if the model were true. If the visual inspection of the data (i.e., of the plot we have labeled $T(y)$) shows patterns that do not generally appear in the replications, $T(y^{\text{rep}})$, then the exploratory analysis has indicated a potential misfit of model to data. This idea was presented in the context of permutation tests by Buja, Cook, and Swayne (1999). Our treatment here can be seen as a generalization to reference distributions based on parametric models.

2.2 EXAMPLE OF PREDICTIVE CHECKING FOR EXPLORATORY AND CONFIRMATORY DATA ANALYSIS

Figure 1 shows an example from a model fit to the delays of capital appeals in state courts, from 1973, when the death penalty was reintroduced in the United States, through 1995 (see Gelman, Liebman, West, and Kiss 2004). The upper-left plot in the figure shows the actual data of delays plotted against the year of sentencing. Delays are variable, but the average delays increase gradually through the mid-1980s and then decrease. However, some of the pattern is an artifact of the data collection: no information was available after 1995, and thus the observed delay for sentences at year t can never exceed $1995 - t$.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

We modeled these data (along with information on the states in which each death sentence arose, and information on later stages of appellate review) with a hierarchical Weibull regression with censoring. The three plots in Figure 1 labeled as replications display random draws y^{rep} from the posterior predictive distribution under the model. There is a clear visual discrepancy between the data and the replications, with the actual data having consistently shorter delays, on average, in the first few years.

At this point, we can move from exploratory to confirmatory analysis and quantify the discrepancy by defining a relevant scalar test statistic $T(y)$ and calculating its p value.

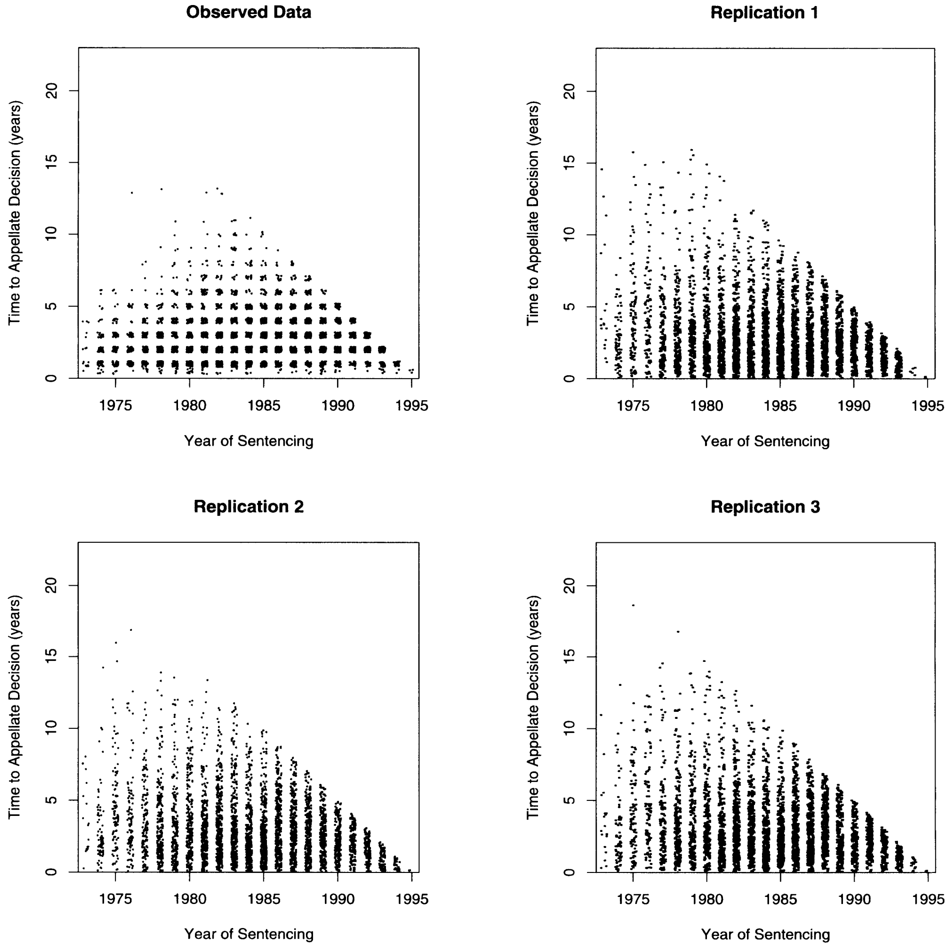


Figure 1. Delays in state appeals court for death penalty cases, plotted versus year of sentencing (jittered to allow individual data points to be visible). The upper-left plot shows the observed data, and the other three plots show replications under the fitted model. (The delays are jittered for the observed data but not for the replications, which have been simulated under a continuous model.) The observed data show a pattern—a steady increase in delay times for the first decade—that is not captured by the replicated datasets. This is an example of how replicated datasets can be used in exploratory data analysis to assess whether an observed pattern is explained by a fitted model.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

For example, suppose we define $T(y)$ to be the number of cases with delays observed to be at least 10 years. The observed value of this test statistic is 26. By comparison, in 200 simulations of from the model, $T(y^{\text{rep}})$ had a median of 79 and a 95% interval of $[61, 98]$. In fact, $T(y^{\text{rep}}) > T(y)$ for all 200 replications, implying a p value of less than .005. The model predicts about three times as many cases with long delays than are actually observed, and that discrepancy is statistically significant.

This example illustrates (a) how a straightforward data display such as the upper-left plot in Figure 1 can be viewed as a model check, and (b) the view that exploratory and confirmatory data analysis are both forms of predictive model checking. We continue with some theoretical discussion and then in Section 3 consider the relation between EDA techniques and invariance properties of replication distributions.

2.3 USING THE POSTERIOR PREDICTIVE FRAMEWORK TO INSPIRE MORE GENERAL FORMS FOR EXPLORATORY GRAPHICS

In the posterior predictive framework, one can define test variables $T(y, \boldsymbol{\theta})$ that depend on unknown parameters as well as data (Gelman, Meng, and Stern 1996). For example, in a regression context, one can plot realized residuals, $y - X\beta$, rather than estimated residuals, $y - X\hat{\beta}$ (Chaloner and Brant 1988). The advantage of realized residuals is that their reference distribution is more simply specified under the model, without needing to correct this distribution for the process of estimating $\hat{\beta}$. For example, consider a nonlinear regression model with independent normal errors and a mix of equality and inequality constraints on the regression coefficients. Depending on the structure of the nonlinear link function and the constraints, the distribution of the estimated residuals can potentially be quite complicated. In contrast, the realized residuals have independent normal distributions under the model.

Realized residuals are particularly useful in hierarchical regressions. Consider a logistic regression model predicting survey respondents' voter preferences y_i from demographic variables X_i and indicators j_i for states, with a second-level linear regression model predicting state coefficients α_j from a state-level linear predictor γW_j including information such as past election results. The distribution of the *realized* state-level residuals $\alpha_j - \gamma W_j$ is conveniently specified by the model, but the *estimated residuals* $\hat{\alpha}_j - \hat{\gamma} W_j$ are more difficult to interpret. For example, a national survey will have fewer respondents from small states, and thus their estimated residuals will be shrunken more toward zero in a hierarchical analysis (see, e.g., Raudenbush and Bryk 2002). A plot of estimated residuals would then show a misleading pattern of higher variances for the small states, even if the underlying variances were equal. Our point here is not that realized residuals are always preferred but rather that the posterior predictive framework is flexible enough to allow them.

Another way that test variables can be generalized is by working with missing and latent data (Gelman et al. in press), so that the dataset that is being replicated can be written as $y = (y_{\text{obs}}, y_{\text{mis}}, y_{\text{latent}})$. Plots of these *completed* datasets y can be easier to interpret, in the sense of having invariance properties that allow simple implicit comparisons to a reference

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

distribution. For a very simple example, consider a model of normally distributed data that are randomly censored in some range. Then it can make sense to view the completed dataset (including imputed values for the censored data) and compare it with a normal distribution. As an exploratory view of the data, this completed-data plot can be more revealing than a plot of the observed data, which must be compared to a nonstandard censored distribution. An example of a completed-data plot is presented near the end of this article.

From an exploratory data analysis perspective, test variables that involve parameters or missing/latent data provide a graphical challenge because the test variable is now itself random and can be represented by several draws from the posterior distribution of completed data and parameters. The completed data can be displayed as multiple graphs—a graphical analogue to the method of multiple imputation for missing data (Rubin 1996).

Comparisons between data and replications can sometimes be sharpened by viewing differences, $D(y, y^{\text{rep}}) = T(y) - T(y^{\text{rep}})$, or more generally, any discrepancy measure that is an antisymmetric function of y and y^{rep}. (In this and subsequent notation, we are implicitly allowing y and y^{rep} to include parameters and missing/latent data.) Predictive simulations of an antisymmetric measure $D(y, y^{\text{rep}})$ can then be compared to the random distribution centered about 0 that would be expected if the observed data actually came from the model (Berkhof, Van Mechelen, and Gelman 2002).

2.4 A Simple Theoretical Example Showing the Potential of Exploratory Data Analysis to Resolve Difficulties in Statistical Modeling

We are used to thinking of exploratory data analysis as an approach to finding unexpected aspects of the data; that is, aspects not captured by an existing model. In addition, exploratory data analysis can reveal modeling problems that could have been anticipated theoretically but were not. As a result, routine use of predictive model comparison can reduce the need for statistical theory. This is related to the idea from the bootstrap literature that simulation can replace mathematical analysis (Efron and Tibshirani 1993).

We illustrate with an example (from Gelman 2003) where an inherent difficulty of estimation is revealed by comparing data to predictive simulations. The example is the fitting of a simple mixture model with unconstrained variances:

$$p(y_i|\mu_1, \mu_2, \sigma_1, \sigma_2) = .5\frac{1}{\sigma_1}\phi\left(\frac{y_i - \mu_1}{\sigma_1}\right) + .5\frac{1}{\sigma_2}\phi\left(\frac{y_i - \mu_2}{\sigma_2}\right), \tag{2.1}$$

where ϕ is the unit normal density function. When fit to data $y_i, i = 1, \dots, n$, the likelihood can be made to approach infinity by setting μ_1 equal to y_i—for any of the data points y_i—and letting σ_1 approach 0. At this limit, the likelihood for y_i approaches infinity, and the likelihoods for the other data points remain finite (because of the second mixture component), so the complete likelihood blows up. This will happen for any data vector y.

Bayesian inference does not necessarily resolve this problem. For example, if a uniform prior distribution is assigned to $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2)$, then the posterior modes will still be

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

at the points where one or another of the σ's approach 0, and these modes in fact contain infinite posterior mass.

But now suppose that we use exploratory data analysis—plotting the data vector y (as a histogram, since the data are unordered and univariate) and comparing to replicated data from the model. Under maximum likelihood, these would be datasets y^{rep} drawn from $p(y^{\text{rep}}|\hat{\boldsymbol{\theta}})$; a Bayesian would use simulations from the posterior predictive distribution, $p(y^{\text{rep}}|\boldsymbol{\theta}, y)$. For this problem, it is not particularly important which method is used. In either case, there are two likely possibilities:

1. At least one of the modes (with their infinite likelihood and infinite posterior mass) is found, in which case each simulated y^{rep} will look like a mixture of a spike at one point and a broad distribution for the other half of the data. (Recall that in this example the model is constrained to give the two modes equal weight, so in any replicated dataset, approximately half the points will fall in each mode.) The misfit of model to data will then be apparent, either from a visual comparison of the histogram of the data y to the histogram of the y^{rep}'s, or using an antisymmetric discrepancy function such as the difference between the histograms of y^{rep} and y. The discrepancy could be summarized by the p value from a numerical discrepancy such as the Kolmogorov-Smirnoff distance between the empirical distributions of y^{rep} and y.

2. Or, the estimation procedure could behave well and fail to find the degenerate modes. In this case, simulated replicated data could look quite similar to the actual data, and no problem will be found. And this would be fine, because the computational procedure is in effect fitting a truncated model that fits the data well.

In either case, exploratory data analysis techniques applied to the fitted model have succeeded in discovering an important problem, if it arose in the estimation. In contrast, a key problem with model-based inference—if exploratory analysis is not performed—is that if an inappropriate model is fit to data, it is possible to end up with highly precise, but wrong, inferences.

3. RELATING METHODS OF EXPLORATORY DATA ANALYSIS TO PROPERTIES OF STATISTICAL MODELS

It has long been noted that exploratory data analysis techniques are particularly effective when they exploit symmetry properties, so that the eye is drawn to patterns violating expected symmetries or repetitions (Tukey 1977; Tufte 1990). Buja et al. (1988) and Buja, Cook, and Swayne (1999) used symmetries to conduct permutation tests, thus constructing reference distributions for visual comparisons. This section uses these ideas to understand exploratory plots in terms of implicit models, and to suggest ways in which model-based tests can be displayed graphically.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

3.1 Theories of Statistical Graphics

One of the frustrating aspects of teaching and practicing statistics is the difficulty of formalizing the rules, if any, for good statistical graphics. As with written language, it takes time to develop a good eye for which graphical displays are appropriate to which data structures, and it is a challenge to identify the "universal grammar" underlying our graphical intuitions (see Wilkinson 1999). Meanwhile, students and researchers untrained in graphical methods often seem to have a horrible tendency toward graphical displays that seem perversely wasteful of data (see Gelman, Pasarica, and Dodhia 2002). For an embarrassing example from our own work, Table 7.5 of Gelman et al. (1995) displays tiny numbers with far too many significant figures. The reader can see little but the widths of the columns of numbers; the implicit comparison is then to columns of equal width, which is not particularly interesting from a substantive perspective in that example.

Many interesting perspectives have been given in recent years evaluating the choices involved in graphical displays (e.g., Ehrenberg 1975; Tukey 1977; Tufte 1983; Cleveland and McGill 1984; Cleveland 1985; and Wilkinson 1999). This work has often taken a psychological perspective, assessing the visual judgments made in reading a table or graph, with reference to psychological models of cognition and visual perception. Empirical studies have compared the understandability of the same data graphed different ways, following the principles and methods of experimental psychology. In parallel with this research have come developments of new graphical methods (examples in static graphical displays include Chambers et al. 1983; Tufte 1990; Wegman 1990; and Cleveland 1993).

Existing theories of statistical graphics—that is, what sorts of graphs are good or bad, and when should different sorts of graphs be used—seem to fall in three categories. First, there are general guidelines that seem to make sense (at least to the proposer of the guidelines), such as minimizing non-data-ink (Tufte 1983) and avoiding pie charts (suggested by many authors). Second, there have been some psychological studies (see Cleveland 1985 and Gelman, Pasarica, and Dodhia 2002, for brief reviews). These are interesting and ultimately essential to statistical communication but somehow they do not seem to tell the whole story. Third, it is sometimes possible to use theoretical arguments for graphical procedures (e.g., in deciding the number of bins in a histogram—see Scott 1999; Wand 1997).

We seek here to formalize statistical graphics in a slightly different way—related to the idea of quantifying information context, but focused on the idea of a graph as an explicit or implicit comparison, as discussed by Buja, Cook, and Swayne (1999) and later, in a Bayesian context, by Gelman (2003). Once we systematically think of graphs as model checking, we can think of ways that a graphical display can take advantage of symmetries in the reference distribution of $T(y^{\text{rep}}, \boldsymbol{\theta})$. Conversely, certain graphical displays can be misleading because they implicitly assume symmetries that are inappropriate to the model being considered.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

3.2 Adapting Graphical Forms to the Structures of Test Statistics

The mathematical structure of test statistics and their reference distributions can be used to set up graphical structures to best allow us to detect discrepancies between model and data. Gelman (2003) laid out the following structure to interpret statistical graphics in terms of implicit models, or conversely to display model summaries graphically. This is related to the discussion of "informal statistical inference" by Buja et al. (1988), who categorized graphical structures for permutation tests.

1. The most basic exploratory graphic is simply a display of an entire dataset (or as much of it as can be conveyed in two dimensions). If we think of this display as a test variable $T(y)$, then alongside it we can compare to displays of several draws of $T(y^{\mathrm{rep}})$ from the reference distribution. Figure 1 gives an example.

As discussed by Tufte (1983), Cleveland (1985), and others, displaying data is not simply a matter of dumping a set of numbers on a page (or a screen). For example, Wainer (2001) and Friendly and Kwan (2002) demonstrated the benefits of ordering data before tabulating or graphing them. From a "complex modeling" perspective, there is an advantage to displays whose reference distributions have invariance properties, as we discuss in point 7 below.

2. Figure 2 shows the perils of attempting to interpret data *without* comparing to a reference distribution—the apparent patterns in these maps can be explained by sampling variation. The counties in the center-west of the country have relatively small populations, hence more variable cancer rates and a greater proportion of the highest and lowest values.

It is not immediately clear how best to perform a posterior predictive check in this example—a model would be required—but it is clear that if the map display is used to detect patterns in the data (i.e., as an exploratory data analysis), then some reference distribution is required. Simply looking at the maps for patterns is, implicitly, a comparison to a reference distribution in which all counties are independently and equally likely to be shaded on the map. Such a reference distribution does not make sense statistically. As discussed by Gelman and Price (1999), other approaches to mapping the data also yield "artifacts," that is, systematic patterns that would be expected under the replication distribution. This is an example where statistical modeling is needed in order to perform a reasonable exploratory data analysis.

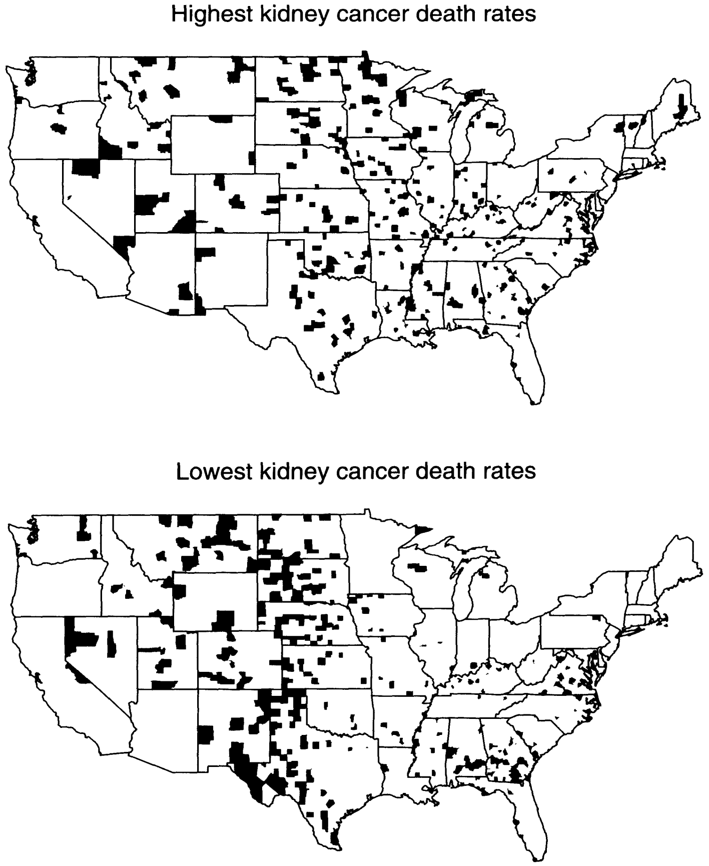
3. If the dataset is large enough, it may have enough internal replication so that the display of a single replicated dataset may be enough to make a clear comparison. Ripley (1988, p. 6) discussed why internal replication is crucial in time series and spatial statistics (where one is often called upon to make inferences from a single sample), and Ripley (1988, chap. 6) presented a striking example of visual comparison of a dataset to simulated data from various models. In this latter example, a spatial model fit to an image fits the second-order statistics (spatial covariance function) of the data essentially perfectly, but a single

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

simulated replicated image y^{rep} from the estimated model looks much different from the data image. These systematic differences motivate the construction of an improved model that fits well, both in its second-order statistics and its overall appearance.

4. At the opposite extreme, if we have a scalar test summary, we can overlay it on a histogram of its simulated reference distribution and thus see both the magnitude of the discrepancy and the p value. Figure 3 presents an example, in this case showing two test statistics that are consistent with replications under the model.

5. A two-dimensional summary can similarly be shown in comparison to a scatterplot, as in Figure 4.



Highest kidney cancer death rates

Lowest kidney cancer death rates

Figure 2. The counties of the United States with the (a) highest and (b) lowest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. Either of these maps appears to show patterns (most of the shaded counties are in the center-west of the country) but they can in fact be explained as artifacts caused by varying sample sizes. (The counties in the center-west of the country have low populations, and so they are more likely to have very high or very low raw rates, just from small-sample variability.) From Gelman and Nolan (2002).

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

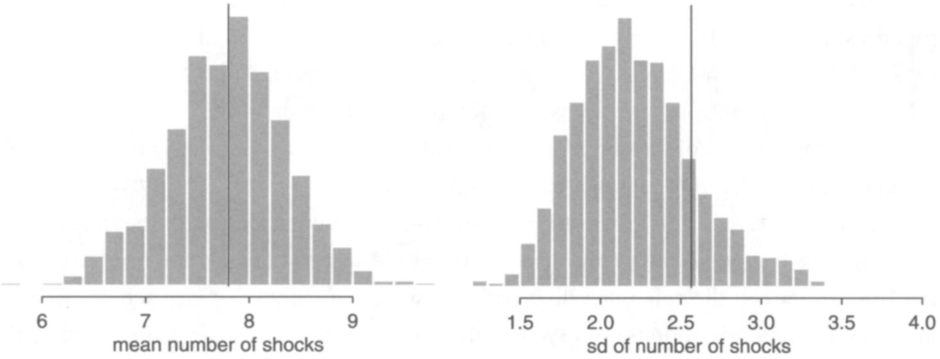


Figure 3. Posterior predictive checks from a stochastic model fit to data of dogs learning to avoid shocks. Each dog was given 25 trials, and the test statistics shown here are the mean and standard deviation, across dogs, of the total number of shocks received. For each plot, the vertical line shows the observed value $T(y)$, and the histogram shows the distribution of the values $T(y^{\text{rep}})$ in simulated replications of the model.

6. A multidimensional summary, $T(y) = (T_1(y), \ldots, T_k(y))$, can be shown as a scatterplot of $T_k(y)$ versus k, in comparison with several scatterplots of $T_k(y^{\text{rep}})$ versus k. But this comparison can be displayed much more compactly using line plots or parallel coordinate plots (Inselberg 1985): a single graph can show the line of $T_k(y)$ versus k in bold, overlaying several lines of $T_k(y^{\text{rep}})$ versus k, each corresponding to a different draw from the reference distribution. Figure 5 presents an example. Wegman (1990) and Miller and Wegman (1991) discussed parallel coordinate plots in detail and in the larger context of displaying multivariate data.

Our suggestion here is to view such plots as model checks, and to apply them to test summaries as well as to raw data.

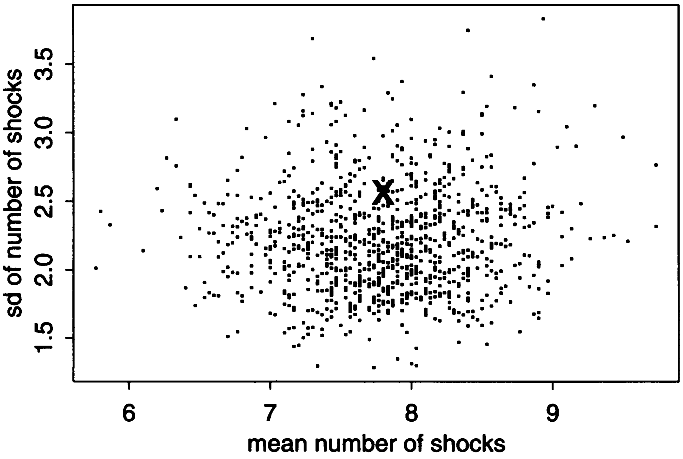
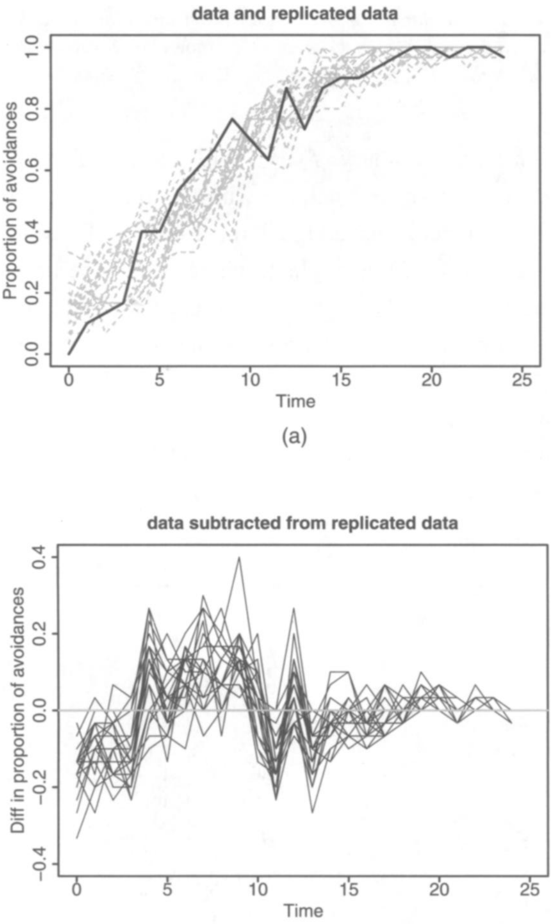


Figure 4. Simultaneous posterior predictive check for the two test statistics shown in Figure 3. The observed value is shown as a bold X on the plot.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

7. Plots can be usefully simplified if the reference distribution has certain invariance properties. For example, consider a binned residual plot of averages of residuals r_i versus predicted values x_i. The range of the predicted values x is divided into K bins, and within each bin k we compute $\bar{x}_k$, the average of the predicted values in the bin, residual in the bin, and $\bar{r}_k$, the average of the corresponding residuals. Figure 6 illustrates with an example of a regression model fit to precinct-level election data.

Under the reference distribution, the residuals are independent, with mean zero, and thus their binned averages also have these properties. As a result, we do not need to display the overlain lines in the reference distribution—since the values $\bar{r}_k^{\rm rep}$ at each point k are independent, no information is conveyed by their joint distribution—the connections of the



(a)

(b)

Figure 5. (a) Plot of the proportion of avoidances among the dogs in the shock-avoidance experiment, as a function of the trial number. The solid line shows the data, and the light lines represent 20 simulated replications from the model. This plot can be seen as a posterior predictive check of a vector test statistic, $T(y)$, compared to replications $T(y^{\rm rep})$. (b) Plots of $T(y^{\rm rep}) - T(y)$ for the 20 simulations of $y^{\rm rep}$. The systematic differences from the horizontal line represent aspects of the data that are not captured by the model.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

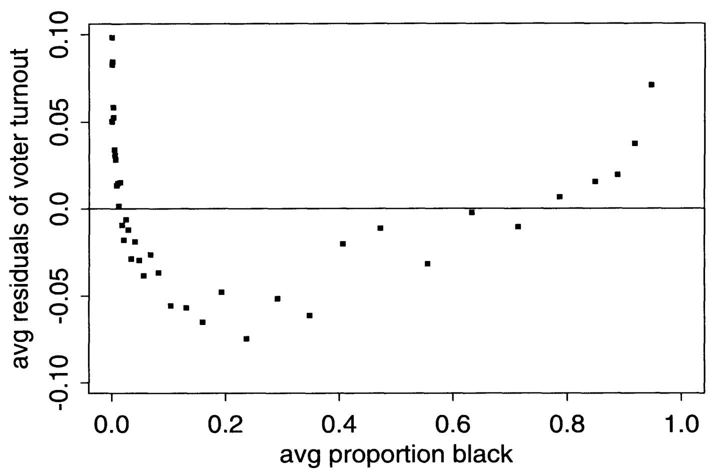


Figure 6. Average residuals $\bar{r}_k$ versus average predicted value $\bar{x}_k$, for regression residuals averaged into 40 bins. The regression was a model of voter turnout on demographics for 5,000 precincts in a New York City mayoral election. Under the model, the binned residuals are independent with mean zero. The systematic discrepancies from zero thus indicate a model misfit, without need for explicit comparison to a reference distribution. From Gelman et al. (2001).

lines, if displayed, would be purely random. As a side benefit, we are also able to remove the lines connecting the dots for the data residuals, because there is no longer a background of replication lines. Instead, the dots are compared to the implicit independence distribution.

The binned residual plot can display more information about the reference distribution by showing pointwise error bounds under the reference distribution, which can be computed via simulation. If the number of points averaged within each bin is reasonably large, the mean residuals $\bar{r}_k$ are approximately normally distributed. We can then display the reference distribution as 95% error bounds. Figure 7 illustrates with binned residuals of pain measurements scores from a nonlinear model in a medical study. Under the model, the residuals are independent with mean zero, and so the systematic pattern of negative residuals at one end and positive residuals at the other indicates a systematic problem.

8. For both of the examples shown in Figures 6 and 7, the model parameters were estimated precisely enough that it was acceptable to display point residuals, $y_i - \mathrm{E}(y_i|\hat{\boldsymbol{\theta}})$. More generally, one could work with realized residuals (see Section 2.3), in which case the residual plots would be random variables. The uncertainty in these plots could be shown by connecting the dots by lines and displaying several overlain, as in Figure 5. That earlier figure is not a residual plot (it graphs replicated data minus observed data) but shares the property that, under the model, its points would be expected to have mean 0.

9. Conversely, confusion can arise when an invariance is wrongly assumed when it is *not* implied by the reference distribution. We have already shown one example with the maps of extreme values in Figure 2.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

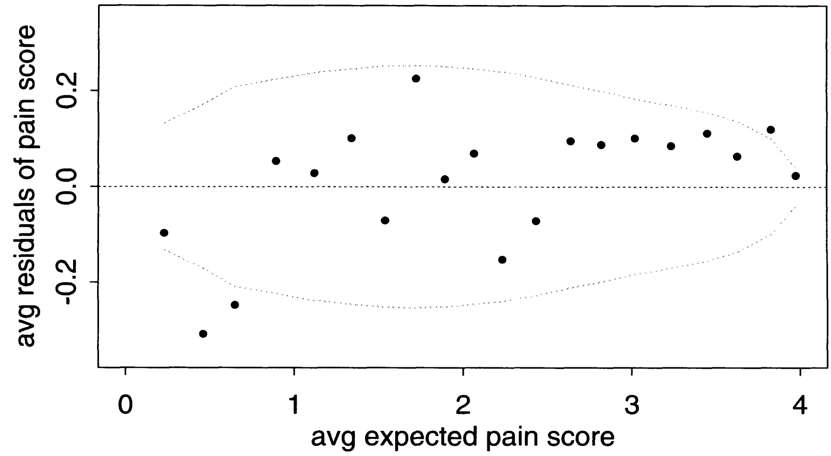


Figure 7. Plot of average residuals versus expected values for a nonlinear model of data from a pain relief experiment, with responses divided into 20 equally sized bins defined by ranges of expected pain scores. The prediction errors are relatively small but with a consistent pattern that low predictions are too low and high predictions are too high. Dotted lines show 95% bounds under the model. Adapted from Gelman and Bois (1997).

For another example, students are often confused about the interpretation of plots of observed data versus expected values, or expected versus observed. In either case, we can go back to the conception of graphs as comparisons to a reference distribution. Under the reference distribution of any model, E(observed|expected) = expected, which implies that the regression line in the plot of observed versus expected should have a slope of 1 (and, of course, the regression line of residuals versus expected should have a slope of 0). The plots of expected versus observed (or residuals versus observed) do *not* in general have any simple reference distribution, and so they are more difficult to interpret.

10. A key principle of exploratory data analysis is to exploit regular structure to display data more effectively (e.g., the "small multiples" of Tufte 1990). The analogy in modeling is hierarchical or multilevel modeling, in which batches of parameters capture variation at different levels. When checking model fit, hierarchical structure can allow us to compare batches of parameters to their reference distribution. In this scenario, the replications correspond to new draws of a batch of parameters.

The top row of Figure 8 shows an example of poor fit (clearly revealed by a single simulation draw of the parameter vectors) from a model in psychometrics (Vansteelandt and Van Mechelen 1998). We can see the lack of fit clearly using the a suspended rootogram, a plot of the difference between the square root of the histogram counts and the square root of the expected counts from the model. Under the model, the histogram counts are Poisson distributed, and the suspended rootogram values should be independent with approximate mean 0 and standard deviation .5 (Tukey 1972). This is an example of a traditional exploratory data analysis plot being used in a modeling setting. We need not display simulations from the reference distribution: the symmetries in the rootogram make this comparison implicit.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

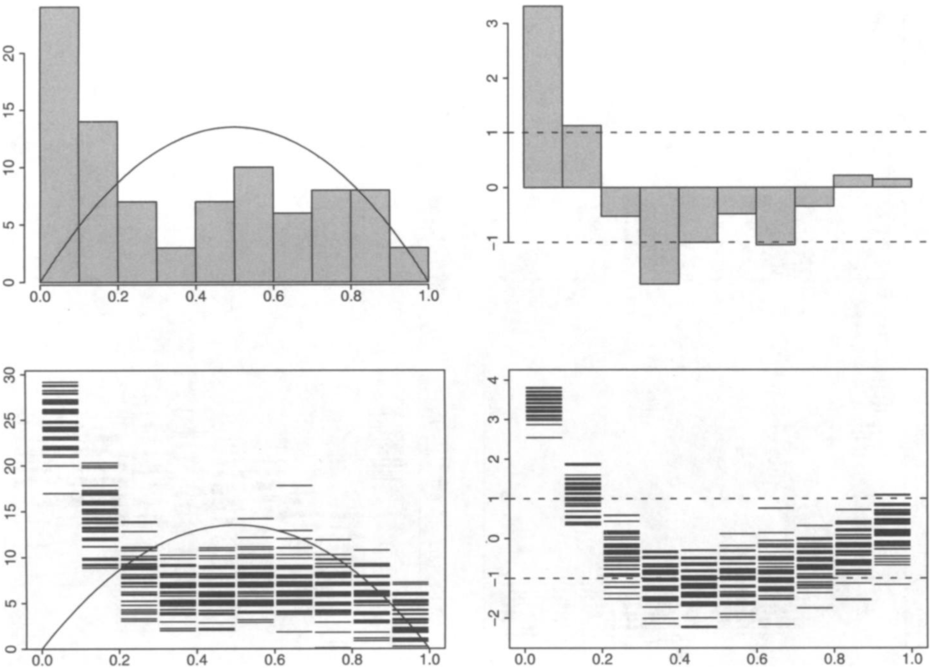


Figure 8. Top row: (a) Histogram of 90 patient parameters θ_j from a single draw from the posterior distribution of a hierarchical model in psychometrics, with the assumed Beta(θ| 2,2) prior density overlain. (b) The suspended rootogram (i.e., square root of the histogram counts, minus the square root of expected counts from the model) clearly shows the misfit. If the model were true, the heights of the suspended rootogram bars would have mean 0 and standard deviation approximately .5; hence the dotted lines at ±1 represent approximate pointwise 95% error bounds (Tukey 1972). Bottom row: the same plots, but showing histograms and suspended rootograms of 50 random draws of the vector θ from its posterior distribution. The counts for the plots in the bottom row have been jittered so the different simulations are distinct.

The bottom row in Figure 8 shows the histogram and the suspended rootogram for 50 random posterior draws of the parameter vector. The shading is omitted from the histograms in order to make the distinct simulations more readable. A large proportion of the 90 parameters are clearly estimated to be near zero, in contradiction to the Beta$(2, 2)$ prior distribution. In the context of the psychometric model, these correspond to individuals that have very low probabilities of having particular psychiatric syndromes.

The misfit in the distribution motivated a new model to be fit—a mixture of beta distributions that allow the individual parameters to have a high probability of being near zero. The new model, and the distribution of the parameters as estimated, appear in Figure 9. The fit is much improved. Our procedure of fitting, checking, and replacing models could be considered a "manual Gibbs sampler," in the sense that the model is being iteratively altered to fit the inferences from the data. This example is typical of hierarchical models in having many parameters with only a moderate amount of data per parameter, thus having the possibility of checking by comparing the histogram of estimated parameters to their prior distribution.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

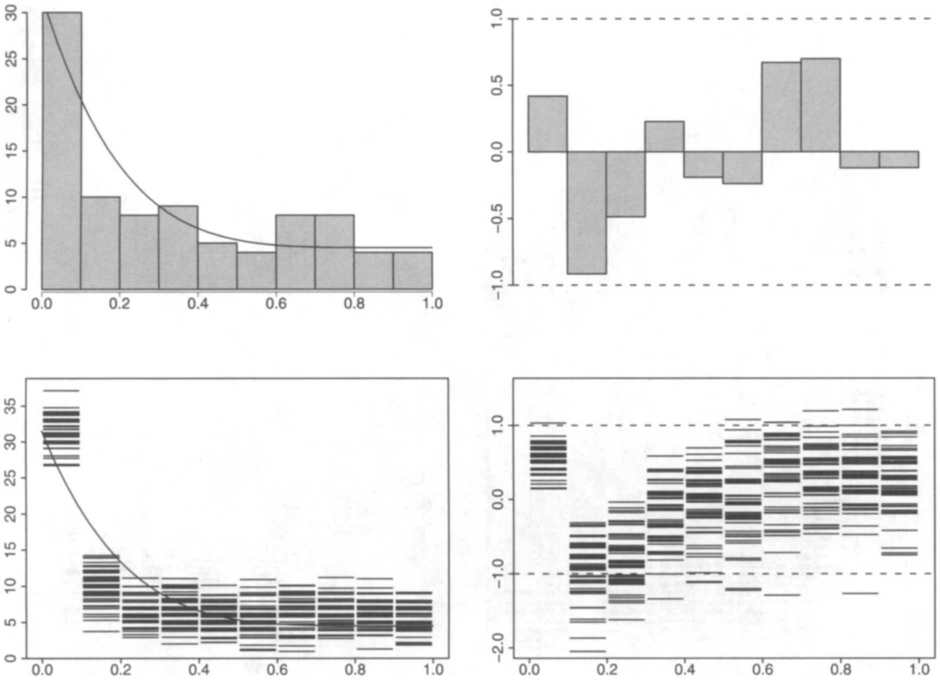


Figure 9. Top row: (a) Histogram of 90 patient parameters $\boldsymbol{\theta}_j$ as estimated from an expanded psychometric model, with the fitted mixture prior distribution overlain. (b) The suspended rootogram shows the simulations to fit the model reasonably well. Compare to Figure 8. Bottom row: the same plots, but showing (jittered) histograms and suspended rootograms of 50 random draws of the vector $\boldsymbol{\theta}$ from its posterior distribution.

11. Finally, in some cases, aspects of a reference distribution are implied, not from symmetries in the model or test statistic, but from external subject-matter knowledge. Figure 10 shows an example from the model of death penalty appeals discussed earlier. The two plots in Figure 10 show simulated completed datasets from two different fitted models. Each plot includes the observed data (as shown in the upper-left plot of Figure 1, p. 760) and a single random imputation for the vector of the missing data—the cases censored because they were still in the review process as of 1995, the endpoint of our study. The left plot in Figure 10 shows a problem with the model, in that the waiting times for cases from early years are bimodal, with a gap between the cases with delays of less than 10 years and those of waits of more than 15 years. Even in the absence of any comparison to a reference distribution, this pattern in the complete data is not plausible. This plot motivated us to go back and clean the data, after which the model was refit, yielding a completed-data plot shown on the right side of Figure 10. (As discussed in point 3 above, the internal replication in this dataset allows us to confidently examine just one completed-data plot as a representative of the entire distribution.)

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

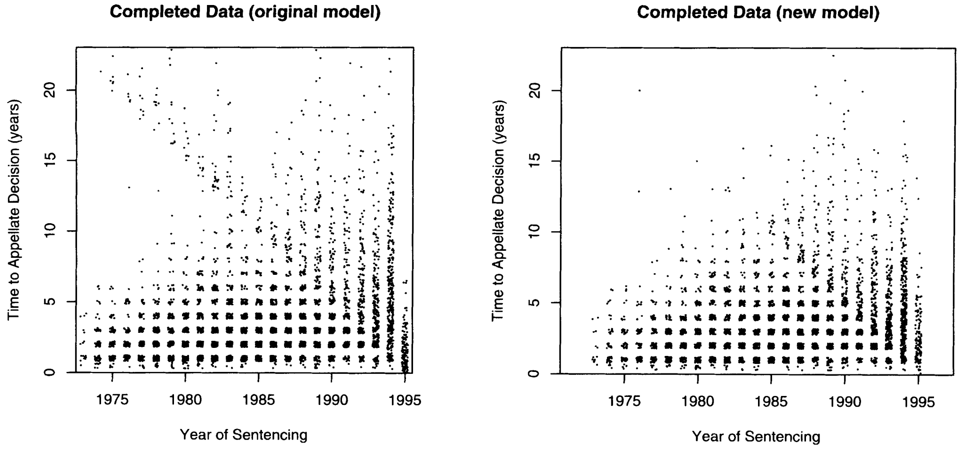


Figure 10. Completed-data plots for the capital appeals data, fit by two different models. In each plot, the observed data are the same points shown in the upper-left plot of Figure 1, and the missing cases (in this example, censored because appellate decisions after 1995 were not in the dataset) have been imputed. The completed data in the left plot do not seem reasonable, in particular as evidenced by the bimodal distribution of delays for the early cases. The data were cleaned, and the new completed-data plot looks more reasonable, although still not perfect (noticeably for the cases from 1995).

3.3 USING GRAPHICAL DESIGN TO HIGHLIGHT STATISTICAL COMPARISONS

Many of the principles of statistical graphics can now be interpreted in light of the dictum that graphs are comparisons to a reference distribution, explicit or implicit. For example, if we follow the dicta to minimize "chartjunk" and unnecessary "ink" (Tufte 1983), we are removing irrelevant information that cannot possibly be informative as a model check. Tufte's (1990) recommendation of "small multiples" (i.e., arrays of several small displays with different data but identical structure) uses the replication in the display to facilitate comparison to the implicit model of no change between the displays. Tukey (1977) recommended rotating plots by 45° so that they can be compared to horizontal lines, and Cleveland (1985) suggested aligning data displays so that lengths, rather than angles, can be compared.

Even very simple recommendations, such as omitting unnecessary grid lines (Tufte 1983) and displaying graphs as square only if the two axes are on a common scale, can be interpreted as removing the opportunity for irrelevant comparisons to focus more clearly on the implicit model checking involved in virtually any data display.

4. INTEGRATING GRAPHICS AND COMPLEX MODELING INTO STATISTICAL PRACTICE

We expect that exploratory data analysis would be more effectively used in modern statistical analyses if it could be routinely implemented as part of software for complex

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

modeling. To some extent this is already done with residual plots in regression models, but there is the potential for much more progress by explicitly defining model-checking plots and replication distributions.

We see room for improvement, leading toward statistical packages with automatic features for simulating replication distributions and performing model checks. We anticipate four challenges:

1. *The computational environment.* We propose to implement replications into statistical computation in two steps. We start with the idea from Bayesian inference (Smith and Roberts 1993) and multiple imputation (Rubin 1996) of representing all unknown quantities by some number L of simulation draws. Thus, a scalar parameter or missing value $\boldsymbol{\theta}$ is represented by a vector of L posterior simulations; a vector of length J is stored as an $L \times J$ matrix of simulations; an uncertain quantity that is a $J \times K$ matrix (e.g., a set of J parameters for each of K groups, or a set of J latent measurements on each of K persons) is stored as an $L \times J \times K$ array; and so forth. This part of our structure is already present in the BUGS package for Bayesian inference (Spiegelhalter, Thomas, Best, and Gilks 1994, 2002) and the MICE software for multiple imputation (Van Buuren and Oudshoom 2000).

Just as the concept of "data frame" expands "data matrix" to allow for more general structures (Chambers and Hastie 1992; Chambers 1998; Billard and Diday 2003), we would like to be able to work with parameters and data directly, with the details of the simulations hidden by the storage structure. Thus, for example, if a data matrix y has some observed and some missing values, we can simply work with y as a matrix, with the understanding that each missing value is represented by a vector of L simulations.

The second part of our proposal is to duplicate the entire "data and parameter space" for the replications. The structure of the replications will depend on the complexity of the model. For a simple nonhierarchical model with parameter vector $\boldsymbol{\theta}$ estimated by maximum likelihood, the replications y^{rep} can be simulated from the model given $\hat{\boldsymbol{\theta}}$. The Bayesian counterpart is posterior predictive checking, where for each posterior simulation of $\boldsymbol{\theta}$ is used to simulate a predictive dataset y^{rep} (Meng 1994).

With missing/latent data, the vector of inclusion indicators I should be added to the model (where, for each potential observation i, $I_i = 1$ if data point y_i is observed), and the replication process creates both y^{rep} and I^{rep} given the parameters estimated from the model; thus simulating a completed underlying dataset as well as indicators for which data would be observed in the replication (Gelman et al. in press). Each of these would be represented by L simulation draws but these would be transparent to the user.

In hierarchical models, the replicated structure can become more complicated (see Gelman, Meng, and Stern 1996). For a model with lower-level parameters $\boldsymbol{\theta}$ and hyperparameters ϕ, it can make sense to simulate replicated parameters $\boldsymbol{\theta}^{\text{rep}}$ along with data y^{rep}. This is similar to the "empirical Bayes" approach in classical statistics (see Morris 1983) in which lower-level parameters are treated as missing data. The implicit replication of parameters is the basis of the model checks in Figures 8–9.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

2. *The replication distribution.* Constructing a replication distribution is analogous to the problem of specifying the prior distribution in a Bayesian analysis (Gelman 2003). It can never be automatic, but standard options will be possible. For example, in a language such as BUGS (Spiegelhalter et al. 1994), replications will have to be defined for all data and parameters in the model, and a simple start would be to choose, for each, the option of resampling it or keeping it the same as with the current inference.

Resampling a parameter might require more modeling effort than keeping a parameter fixed, which is as it should be, because ultimately it defines an assumption about how the model will be used. For example, in analyzing sample survey data, the sample size n could be drawn from a Poisson distribution (with a model $n \sim \text{Poisson}(\lambda)$ and a replication distribution $n^{\text{rep}} \sim \text{Poisson}(\lambda)$) or simply fixed (so that $n^{\text{rep}} \equiv n$). The choice would depend on the design of the study, and more complicated replication models could be appropriate, for example if analyzing data collected sequentially.

In more conventional software such as SAS, SPSS, and Stata, standard models of replications can be assigned along with the currently implemented menus of models for inference. Options can be created to allow different replication structures, in the same way that model classes such as generalized linear models have options for assigning different link functions.

3. *The test variables.* In the spirit of exploratory data analysis, it would be appropriate to look at many data summaries rather than trying to select a single test statistic. The computational environment should allow the user to simply identify a data summary as a "test variable" and then plot it, comparing to its reference distribution. These would include plots of the observed dataset, as in Figure 1 (p. 760), and completed data, as in Figure 10. In addition, it would make sense to have a set of automatically chosen test variables, going beyond the current defaults such as residual and quantile plots for continuous regression models and their analogies in discrete-data regressions (e.g., Atkinson 1981, 1985; Landwehr, Pregibon, and Shoemaker 1984; Gelman et al. 2000).

In complex models (the subject of this article), test variables can be constructed using structure in the model or data. For example, data and residuals can be averaged at the group level and plotted versus group-level predictors (see Figure 5, p. 768), and vectors of exchangeable variables at any level of the model can be displayed as histograms (as in Figures 8–9). More complicated cross-level structures, such as occur in latent class models, could also be plotted. This is still an open area, but we envision that structure in the model could define default structures in the test variables, generalizing the ideas of Tukey (1972, 1977) on two-way plots.

The parameterization under which the model is set up would then affect the way the model is tested, which makes sense because we often understand models in terms of their parameters and variables. For example, consider two equivalent ways of expressing a regression in BUGS. If we write,

```
y[i] ~ dnorm (y.hat[i], tau)
y.hat[i] <- a + b*x[i],
```

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

then the vectors y, y.hat, and x are each vectors that could automatically be included in the model checks; for example, as separate histograms and as a plot of y versus y.hat. But if we write the model as

```
y[i] = a + b*x[i] + e[i]
e[i] ~ dnorm (0, tau),
```

then an automatic model-checking program would display the vector of errors e as well, which might be useful for exploratory analysis.

4. *Graphical display.* Finally, the display of the test variables would depend on the dimensionality, structure, symmetry properties, and internal replication of their reference distributions, as discussed in Section 3. For example, a vector of exchangeable parameters can be displayed as histograms or quantile plots, and any $2 \times n$ matrix structures in the data or model can be displayed as scatterplots. More complicated structures would require some default means of display, perhaps parallel plots with observed or realized data on the left and replicates on the right, and the entire system would need to be linked to a language such as R (R Project 2000) or S-Plus (Mathsoft 2000) that allows programmable graphics.

There is need for further work in this area in various directions. The statistical theory of model checking has now been set up in a Bayesian framework, but there are various open problems, especially in understanding the sampling properties of predictive p values (Meng 1994; Bayarri and Berger 1998) and simultaneous inference (Buja and Rolke 2003). Theory has also been developed for the related method of cross-validation (Stone 1994; Gelfand, Dey, and Chang 1992). In the direction of statistical graphics, it is not always clear how to best display a given data structure when the goal is comparison to a reference distribution with given invariance properties. The kind of subtle structures that we look for in exploratory data analysis involve many simultaneous visual comparisons, and the problem becomes potentially more difficult once dynamic graphics are allowed (see, e.g., Buja, Cook, and Swayne 1996).

However, we believe that the current state of statistical theory and graphics allows for effective exploratory data analysis in a wide range of complex modeling situations. As models and data structures become more complex, we anticipate corresponding developments in exploratory data displays, with the "replicated data" idea being used to interpret data displays in models with structure, high dimensionality, and missing and latent data.

ACKNOWLEDGMENTS

We thank Yves Chretien for the analysis of the death penalty data, Andreas Buja, Hans Berkhof, David Scott, Francis Tuerlinckx, Iven Van Mechelen, and reviewers for helpful discussions, and the associate editor for suggesting the rootograms in Figures 8–9. This work was supported in part by Fellowship F/96/9 and Grant OT/96/10 of the Research Council of Katholieke Universiteit Leuven and grants SBR-9708424, SES-9987748, SES-0318115, and Young Investigator Award DAMS-9457824 of the U.S. National Science Foundation.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

[Received July 2002. Revised July 2003.]

REFERENCES

Atkinson, A. C. (1981), "Two Graphical Displays for Outlying and Influential Observations in Regression," *Biometrika*, 68, 13–20.

——— (1985), *Plots, Transformations, and Regression*, Cambridge, MA: Oxford University Press.

Bayarri, M. J., and Berger, J. (1998), "Quantifying Surprise in the Data and Model Verification," in *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York: Oxford University Press, pp. 53–82.

Berkhof, J., Van Mechelen, I., and Gelman, A. (2002), "Posterior Predictive Checking Using Antisymmetric Discrepancy Functions," technical report.

Billard, L., and Diday, E. (2003), "From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis," *Journal of the American Statistical Association*, 98, 470–487.

Buja, A., Asimov, D., Hurley, C., and McDonald, J. A. (1988), "Elements of a Viewing Pipeline for Data Analysis," in *Dynamic Graphics for Statistics*, ed. W. S. Cleveland and M. E. McGill, Belmont, CA: Wadsworth, pp. 277–308.

Buja, A., Cook, D., and Swayne, D. (1996), "Interactive High-Dimensional Data Visualization," *Journal of Computational and Graphical Statistics*, 5, 78–99.

——— (1999), "Inference for Data Visualization," talk given at Joint Statistical Meetings [on-line]. Available at www.research.att.com/~andreas/#dataviz.

Buja, A., and Rolke, W. (2003), "Calibration for Simultaneity: (Re)sampling Methods for Simultaneous Inference with Applications to Function Estimation and Functional Data," technical report, Department of Statistics, University of Pennsylvania.

Bush, R. R., and Mosteller, F. (1955), *Stochastic Models for Learning* (chap. 11), New York: Wiley.

Carlin, B. P., and Louis, T. A. (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, London: Chapman and Hall.

Chaloner, K., and Brant, R. (1988), "A Bayesian Approach to Outlier Detection and Residual Analysis," *Biometrika*, 75, 651–659.

Chambers, J. M. (1998), *Programming With Data*, New York: Springer-Verlag.

Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), *Graphical Methods for Data Analysis*. Pacific Grove, CA: Wadsworth.

Chambers, J. M., and Hastie, T. J. (1992), *Statistical Models in S*, London: Chapman and Hall.

Cleveland, W. S. (1985), *The Elements of Graphing Data*, Monterey, CA: Wadsworth.

——— (1993), *Envisioning Information*, Summit, NJ: Hobart Press.

Cleveland, W. S., and McGill, R. (1984), "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods," *Journal of the American Statistical Association*, 79, 531–554.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, New York: Wiley.

Efron, B., and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, London: Chapman and Hall.

Ehrenberg, A. S. C. (1975), *Data Reduction: Analysing and Interpreting Statistical Data*, New York: Wiley.

Finch, P. D. (1979), "Description and Analogy in the Practice of Statistics," *Biometrika*, 66, 195–208.

Friendly, M., and Kwan, E. (2002), "Effect Ordering for Data Displays," *Computational Statistics and Data Analysis*, 43, 509–539.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

Gelfand, A. E., Dey, D. K., and Chang, H. (1992), "Model Determination Using Predictive Distributions with Implementation via Sampling-Based Methods" (with discussion), in *Bayesian Statistics 4*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, New York: Oxford University Press, pp. 147–167.

Gelman, A. (2003), "A Bayesian Formulation of Exploratory Data Analysis and Goodness-of-Fit Testing," *International Statistical Review.*

Gelman, A., Ansolabehere, S., Price, P. N., Park, D. K., and Minnite, L. C. (2001), "Models, Assumptions, and Model Checking in Ecological Regressions," *Journal of the Royal Statistical Society*, Ser. A, 164, 101–118.

Gelman, A., and Bois, F. Y. (1997), Discussion of "Analysis of Non-randomly Censored Ordered Categorical Longitudinal Data from Analgesic Trials," by L. B. Sheiner, S. L. Beal, and A. Dunne, *Journal of the American Statistical Association.*

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995), *Bayesian Data Analysis* (1st ed.), London: Chapman and Hall.

Gelman, A., Goegebeur, Y., Tuerlinckx, F., and Van Mechelen, I. (2000), "Diagnostic Checks for Discrete-Data Regression Models using Posterior Predictive Simulations," *Applied Statistics*, 49, 247–268.

Gelman, A., Liebman, J., West, V., and Kiss, A. (2004), "A Broken System: The Persistent Patterns of Reversals of Death Sentences in the United States," *Journal of Empirical Legal Studies*, 1, 209–261.

Gelman, A., Meng, X. L., and Stern, H. S. (1996), "Posterior Predictive Assessment of Model Fitness via Realized Discrepancies" (with discussion), *Statistica Sinica*, 6, 733–807.

Gelman, A., and Nolan, D. (2002), *Teaching Statistics: A Bag of Tricks*, Cambridge, MA: Oxford University Press.

Gelman, A., Pasarica, C., and Dodhia, R. (2002), "Let's Practice What We Preach: Turning Tables into Graphs," *The American Statistician*, 56, 121–130.

Gelman, A., and Price, P. N. (1999), "All Maps of Parameter Estimates are Misleading," *Statistics in Medicine*, 18, 3221–3224.

Gelman, A., Van Mechelen, I., Verbeke, G., Heitjan, D. F., and Meulders, M. (in press), "Multiple Imputation for Model Checking: Completed-Data Plots with Missing and Latent Data," *Biometrics.*

Guttman, I. (1967), "The Use of the Concept of a Future Observation in Goodness-of-Fit Problems," *Journal of the Royal Statistical Society*, Ser. B, 29, 83–100.

Hastie, T., Tibshirani, R., and Friedman, J. (2002), *The Elements of Statistical Learning*, New York: Springer-Verlag.

Inselberg, A. (1985), "The Plane with Parallel Coordinates," *The Visual Computer*, 1, 69–91.

Landwehr, J. M., Pregibon, D., and Shoemaker, A. C. (1984), "Graphical Methods for Assessing Logistic Regression Models" (with discussion), *Journal of the American Statistical Association*, 79, 61–71.

Mathsoft (2000), S-Plus, www.splus.com/.

Meng, X. L. (1994), "Posterior Predictive p Values," *The Annals of Statistics*, 22, 1142–1160.

Miller, J. J., and Wegman, E. J. (1991), "Construction of Line Densities for Parallel Coordinate Plots," in *Computing and Graphics in Statistics*, eds. A. Buja and P. Tukey, New York: Springer-Verlag, pp. 219–230.

Morris, C. (1983), "Parametric Empirical Bayes Inference: Theory and Applications" (with discussion), *Journal of the American Statistical Association*, 78, 47–65.

R Project (2000), "The R Project for Statistical Computing," [on-line], www.r-project.org/.

Raudenbush, S. W., and Bryk, A. S. (2002), *Hierarchical Linear Models* (2nd ed.), Thousand Oaks, CA: Sage.

Ripley, B. D. (1988), *Statistical Inference for Spatial Processes*, New York: Cambridge University Press.

Rubin, D. B. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172.

——— (1996), "Multiple Imputation After 18+ Years" (with discussion), *Journal of the American Statistical Association*, 91, 473–520.

Scott, D. W. (1979), "On Optimal and Data-Based Histograms," *Biometrika*, 66, 605–610.

Sheiner, L. B., Beal, S. L., and Dunne, A. (1997), "Analysis of Non-randomly Censored Ordered Categorical Longitudinal Data from Analgesic Trials" (with discussion), *Journal of the American Statistical Association.*

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

Smith, A. F. M., and Roberts, G. O. (1993), "Bayesian Computation via the Gibbs Sampler and Related Markov Chain Monte Carlo Methods" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 55, 3–102.

Spiegelhalter, D., Thomas, A., Best, N., and Gilks, W. (1994, 2002), "BUGS: Bayesian Inference Using Gibbs Sampling," MRC Biostatistics Unit, Cambridge, England, [on-line], www.mrc-bsu.cam.ac.uk/bugs/.

Stigler, S. M. (1986), *The History of Statistics*, Cambridge, MA: Harvard University Press.

Stone, M. (1974), "Cross-Validatory Choice and Assessment of Statistical Predictions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 36, 111–147.

Tufte, E. R. (1983), *The Visual Display of Quantitative Information*, Cheshire, CT: Graphics Press.

——— (1990), *Envisioning Information*, Cheshire, CT: Graphics Press.

Tukey, J. W. (1972), "Some Graphic and Semigraphic Displays," in *Statistical Papers in Honor of George W. Snedecor*, ed. T. A. Bancroft, Ames, IA: Iowa State University Press.

——— (1977), *Exploratory Data Analysis*, New York: Addison-Wesley.

Van Buuren, S., and Oudshoom, C. G. M. (2000), "MICE: Multivariate Imputation by Chained Equations," web.inter.nl.net/users/S.van.Buuren/mi/.

Vansteelandt, K., and Van Mechelen, I. (1998), "Individual Differences in Situation-Behavior Profiles: A Triple Typology Model," *Journal of Personality and Social Psychology*, 3, 751–765.

Wainer, H. (1997), *Visual Revelations*, New York: Springer-Verlag.

——— (2001), "Order in the Court," *Chance*, 14, (2), 43–46.

Wand, M. P. (1997), "Data-Based Choice of Histogram Bin Width," *Statistical Computing and Graphics*, 51, 59–64.

Wegman, E. J. (1990), "Hyperdimensional Data Analysis Using Parallel Coordinates," *Journal of the American Statistical Association*, 85, 664–675.

Wilk, M., and Gnanadesikan, R. (1968), "Probability Plotting Methods for the Analysis of Data," *Biometrika*, 55, 1–17.

Wilkinson, L. (1999), *The Grammar of Graphics*, New York: Springer-Verlag.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

Discussion

Andreas Buja

Gelman's article is a thought-provoking mix of opinions and creative methodology. I agree with Gelman that the disjunction of models and exploratory data analysis (EDA) in mainstream statistics is unsound. The assumption that models belong exclusively to confirmatory data analysis (CDA) is easily shown to be wrong in practice. Many uses of models, in particular model selection, are exploratory, and they usually forfeit the possibility of strict CDA. Gelman's suggestion that EDA be based on mainstream modeling practices is long overdue.

1. DOES BAYES HAVE AN EDGE?

Gelman is a Bayesian, but he gives a nod to non-Bayesian alternatives of creating reference distributions: permutation tests, bootstrap, and cross-validation (Section 2.1). But then he returns quickly to the Bayesian fold and discusses all further examples in terms of the "posterior predictive framework." It should be stated, however, that most of his treatments of examples have frequentist analogues, so we see very little unique advantage in the Bayesian version of model-based EDA. It would help us agnostics to hear what particular powers the predictive posterior framework brings to bear in EDA, other than aesthetic ones (the Bayesian framework is certainly pretty).

2. HONEST INFERENCE FOR DATA VISUALIZATION

Gelman references our early approaches to visual inference (Buja, Asimov, Hurley, and McDonald 1988, sec. 5; Swayne, Cook, and Buja 1998). More recently, we (Cook and I) became interested in the question of how to add some inferential validity at least to parts of EDA. In various talks (such as www-stat.wharton.upenn.edu/~buja/PAPERS/jsm99.ps.gz) we have discussed the possibility of a protocol that would establish for a given plot whether

Andreas Buja is Professor, Statistics Department, The Wharton School, University of Pennsylvaniz, 471 Huntsman Hall, Philadelphia, PA 19104.

©2004 American Statistical Association, Institute of Mathematical Statistics, and Interface Foundation of North America

Journal of Computational and Graphical Statistics, Volume 13, Number 4, Pages 780–784
DOI: 10.1198/106186004X12281

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

the visually perceived structure is "real" or not. Part of the protocol is to hide a plot of the actual data among a large number of plots of artificial data simulated under a null hypothesis. We give an example in terms of a visual permutation test. In the spirit of this protocol, we suggest that Gelman's Figure 1 be redone as follows: instead of separating the actual data in the top left of the page, insert them randomly among many simulated data. Spotting the actual data would then amount to valid inference. In the case of Figure 1, it is quite clear that the protocol is not needed due to the striking nature of the data artifacts, but as a matter of principle, it would be a partial answer to the often-heard suspicion that extensive data snooping is prone to over-interpretation of the data.

3. SIMULTANEOUS INFERENCE, EDA, AND BAYES

No reminder should be needed that simultaneous inference is of interest to Bayesians, too. Ten statements "*parameter θ_i is in the interval I_i with posterior probability .90*" ($i = 1, \ldots 10$) are to be distinguished from the one statement "*the ten parameters θ_i are simultaneously in the intervals I_i with posterior probability .90 ($i = 1, \ldots 10$).*"

Simultaneous inference is a general concern in EDA because of the potentially large number of statistics considered in the course of a data analysis. It is common to plot, for example, confidence bands around smooths that may amount to confidence intervals for 100 or more statistics. The bands are almost always pointwise, but it is possible to widen the bands suitably to provide simultaneous coverage. Calibration can be achieved in simulations whenever a unique reference distribution is available. Buja and Roelke (2003) have pointed out the generality of achieving simultaneous coverage in a wide variety of inference approaches.

4. BOOTSTRAP VERSUS POSTERIORS: SINGLE VERSUS MULTIPLE PARAMETERS

There is an important difference between Bayesian posterior-based and frequentist bootstrap-based sampling. Here are the two procedures:

- Bayes: repeatedly sample parameter sets $\boldsymbol{\theta}$ from the posterior and sample one artifial dataset y^{rep} from $p(y|\boldsymbol{\theta})$ for each $\boldsymbol{\theta}$.
- Parametric bootstrap: obtain one estimated parameter set $\hat{\boldsymbol{\theta}}$ and repeatedly sample artifial datasets y^{rep} from $p(y|\hat{\boldsymbol{\theta}})$.

Actually, the difference has not so much to do with Bayes or bootstrap: one can easily conceive a Bayes version that sees only one set of parameter estimates by using a posterior mode or mean instead of sampling from the posterior. The important difference is in the type of variability the two approaches see:

- *Between-parameter variation*: Posterior sampling shows variation due to differences between parameter sets that are likely given the observations. That is, one sees variation due to posterior spread.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

- *Within-parameter variation*: Parametric bootstrap sampling shows variation for one single parameter set.

Which one is more desirable? We postpone this question for the moment. First we note that one can quite easily make both approaches see both types of variation:

- Bayes: Not only sample repeatedly parameter sets $\boldsymbol{\theta}$ from the posterior, but sample repeatedly artificial datasets y^{rep} for each $\boldsymbol{\theta}$.
- Bootstrap: Requires a two-stage sampling approach.
  Stage 1. Draw the usual bootstrap samples from the actual data, with nonparametric or parametric bootstrap. From each bootstrap sample, obtain a set of bootstrap parameter estimates, one set of estimates per sample.
  Stage 2. Repeatedly draw parametric bootstrap samples from the bootstrap parameter estimates of the first stage.

In either case, one obtains a nested set of artificial datasets, where the multiple parameter sets define the nesting:

$$y \longrightarrow \begin{cases} \theta_{(1)} \longrightarrow \left\{ y^{\text{rep}}_{(1)1},\ y^{\text{rep}}_{(1)2},\ y^{\text{rep}}_{(1)3},\ \ldots \right. \\ \\ \theta_{(2)} \longrightarrow \left\{ y^{\text{rep}}_{(2)1},\ y^{\text{rep}}_{(2)2},\ y^{\text{rep}}_{(2)3},\ \ldots \right. \\ \\ \theta_{(3)} \longrightarrow \left\{ y^{\text{rep}}_{(3)1},\ y^{\text{rep}}_{(3)2},\ y^{\text{rep}}_{(3)3},\ \ldots \right. \\ \\ \ldots \end{cases}$$

It should now be possible to attribute variation to either source: differences between parameter values, and differences within a given parameter value. We can gain some insight into the roles of the two types of variation by analyzing Gelman's theoretical example of a two-component mixture model.

5. WHAT DISCREPANCY MEASURE?

The two-component mixture model of Section 2.4 illustrates potential estimation problems that require diagnostics. Estimation is likely to produce degenerate estimates in which one component is a spike at one of the observations. Such degenerate estimates are easy enough to spot by the human eye, but how can they be detected in terms of a test? Humans have prior insight that enables them to tell an unreasonable estimate when they see one. They easily notice if replicated data consistently show 50% ties but the actual data do not. Not so a computer, unless it is programmed with a test statistic that measures spiky-ness.

In item "1" of Section 2.4, Gelman describes the situation correctly: "The misfit of model to data will then be apparent, either from a visual comparison of the histogram of the data y to the histogram of the y^{rep}'s,. . . ." He continues by describing two automatic ways

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

of reconstructing the human eyes' performance. He seems to think that the two ways are identical, whereas the first way is correct, while the second is not, but can be fixed. Here is an analysis of Gelman's text:

- He continues by saying: "... or using an antisymmetric discrepancy function such as the difference between the histograms of y^{rep} and y." True: assuming identical bin locations for both, the differences between histogram heights constitute a family of discrepancy measures $T_i(y) - T_i(y^{\text{rep}})$, where i indexes the bins. If one chooses the bin containing the spike of y^{rep}, then the bin height $T_i(y^{\text{rep}})$ of the artificial data will be much greater than the bin height $T_i(y)$ of the real data. Therefore, $T_i(y)$ will be shown as unlikely small by falling in the extreme left tail of the $T_i(y^{\text{rep}})$ values. The remaining problem is that we do not know the spike location beforehand, but this can be remedied with simultaneous coverage for all bins (Buja and Rolke 2003). In summary, simultaneous inference across all bin heights will show the histogram of y to be significantly different from the histograms of the y^{rep}'s.

- Curiously, Gelman finishes his discussion with a throw-away remark that is incorrect: "The discrepancy could be summarized by the p value from a numerical discrepancy such as the Kolmogorov-Smirnoff distance between the empirical distributions of y^{rep} and y." This recipe does in fact not work. A distance between actual and artificial data is not a discrepancy measure because it is symmetric rather than antisymmetric. We do not know how the fitted model could be rejected knowing the distribution of $\text{dist}(y^{\text{rep}}, y)$ for fixed y. Recall how discrepancy measures $T(y) - T(y^{\text{rep}})$ work: they suggest rejection of the model if the replication distribution with regard to y^{rep} is mostly to the left of zero. By comparison, we cannot say anything about the replication distribution of $\text{dist}(y^{\text{rep}}, y)$.

  Is there a solution? There is, but it involves two idependent copies of replication data, y^{rep} and y'^{rep}. The idea is to compare $\text{dist}(y, y^{\text{rep}})$ with $\text{dist}(y'^{\text{rep}}, y^{\text{rep}})$: if the model does not fit well, then y is quite different from y^{rep}, more different than most y'^{rep}. Hence we may define $T(y) = \text{E}_{y'^{\text{rep}}}\,\text{dist}(y, y'^{\text{rep}})$, and $T(y) - T(y^{\text{rep}})$ will be a discrepancy measure that responds to spikes: y^{rep} and y'^{rep} both have spikes, but y does not, which sets y apart.

I find both approaches intriguing because they have the potential of being universal diagnostics for model fit. They are very general ideas:

- Map a dataset y to a collection of test statistics $T_i(y)$, such as the heights of histogram bins, and apply simultaneous inference with regard to the reference distribution $T_i(y^{\text{rep}})$.

- Compare y and y^{rep} in relation to y'^{rep} by using $T(y) = \text{E}_{y'^{\text{rep}}}\text{dist}(y, y'^{\text{rep}})$. Apply inference with regard to the reference distribution $T(y^{\text{rep}})$. The function dist() is not necessarily a distance measure; it could be any two-sample test statistic, including Friedman and Rafky's (1981) minimal spanning tree statistics, which have excellent performance in high dimensions.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

6. VARIATION BETWEEN PARAMETER VALUES OR WITHIN?

In the previous section we made an implicit assumption in analyzing Gelman's mixture example: that the spike was in a fixed location. The spike location is implicit in the parameter set $\boldsymbol{\theta} = (\mu_1, \sigma_1, \mu_2, \sigma_2)$: if σ_i is near zero, then μ_i is the spike location. The requirement of a fixed spike location seems to point to the requirement of a fixed parameter set $\boldsymbol{\theta}$, which in turn seems to exclude between-parameter variation. We realize now that the analysis was targeted at a particular $\boldsymbol{\theta}$ and detection of model misfit between the observed y and this particular model distribution $p(y^{\text{rep}}|\boldsymbol{\theta})$. If multiple $\boldsymbol{\theta}$'s came into play, as they would in posterior sampling, multiple spikes would most likely deteriorate the power of detection of misfit: 50 replicates y^{rep} with the same spike will be more powerful evidence than 50 replicates y^{rep} with 50 different spikes. This leads to the (to me) unexpected conclusion that between-parameter variation is not always desirable, in particular when there are qualitative differences between fitted models, such as differing spike locations.

This raises the question as to the value of between-parameter variation. With the two-component mixture model in mind, one could conjecture that multiple $\boldsymbol{\theta}$'s could give information about the prevalence of degenerate fits. Assuming computational feasibility, one could conceive of a within-parameter analysis for each parameter, indicating for a fixed $\theta_{(i)}$ whether its replicates $y^{\text{rep}}_{(i)j}$ $(j = 1, 2, \ldots)$ are incompatible with y. Subsequent between-parameter analysis will show how often this is the case. In the end, one has a quantitative assessment of the degeneracy problem for this model.

7. CONCLUSIONS

I thank Andrew Gelman for a thought-provoking article. We have independently arrived at similar conclusions on several issues, but his interpretation of EDA as model-checking with artificial replications from a model is the clearest formulation we have seen yet. The clarity of this program could establish EDA as a mainstream research activity, as opposed to an idiosyncratic bag of tricks. This program also demonstrates that EDA is not a preliminary stage of data analysis; it is woven into all stages of data analysis.

REFERENCES

Buja, A., and Rolke, W. (2003), "Simultaneous Inference with Applications to Function Estimation and Functional Data," preprint available on-line at www-stat.wharton.upenn.edu/~buja/PAPER/paper-sim.ps.gz.

Friedman, J. H., and Rafsky, L. C. (1981), "Graphics for the Multivariate Two-Sample Problem," *Journal of the American Statistical Association*, 76, 277–287.

Swayne, D. F., Cook, D., and Buja, A. (1998), "XGobi: Interactive Dynamic Data Visualization in the X Windows System," *Journal of Computational and Graphical Statistics*, 7, 113–130.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

Rejoinder

Andrew Gelman

I appreciate Buja's generous comments and will briefly clarify some issues regarding the role of data visualization in model checking, and the relevance of Bayesian inference to model checking.

1. DATA VISUALIZATION

My article presents four general kinds of model-based graphical diagnostics:

1. Displays of raw data, with a model used to simulate the reference distribution of the displays (as in Figure 1). Conversely, Figure 2 illustrates the difficulty, in general, of interpreting data displays *without* a comparison to a reference distribution.
2. Similar displays of tests of lower-dimensional data summaries (e.g., scalars in Figure 3, and vectors in Figures 4 and 5(a)) compared to simulations from fitted model—again, with discrepancies from the simulations illustrating aspects of the data not explained by the model.
3. Residual plots—put more generally, graphs of differences between data and fitted model which, if the model is true, should follow distributions with invariance properties such as independence and zero mean. Figures 5(b), 6, and 7 illustrate how violations of these invariance properties are visibly apparent without the need for explicit comparisons to simulated replications.
4. Displays of latent data/parameters (as in Figures 8 and 9) or of completed datasets (i.e., combinations of observed and missing or latent data, as in Figure 10).

In methods 1 and 2 above, the displays and test summaries are functions of data alone, with the role of the model being to define a reference distribution for comparison. Thus, if a series of models is developed for a single dataset, one would hope to see the replications looking more and more like the actual data. An attractive example appears in Ripley (1988, chap. 6).

©2004 American Statistical Association, Institute of Mathematical Statistics,
and Interface Foundation of North America

Journal of Computational and Graphical Statistics, Volume 13, Number 4, Pages 785–787
DOI: 10.1198/106186004X12317

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

In methods 3 and 4, the model is used to construct the display, and as improved models are fit to a single dataset, the plots should look increasingly "reasonable" in the sense of being consistent with invariances in the model and with outside knowledge of what the completed dataset should look like.

In all these approaches, the question arises in application of how extreme are the departures from the model. Buja's visual permutation test and simultaneous inference bands are important ideas that go beyond simple display of reference distributions to more systematic or focused comparisons. The visual permutation test seems particularly promising for automatic computer implementations of predictive tests.

Data visualization ideas—from the classical methods of Tukey, Cleveland, and others, to the more recent dynamic approaches of Buja, Cook, and their collaborators—should be useful in understanding model fit by bringing data, residuals, and latent data closer to the user, and in allowing visual comparisons to ever-more-complicated reference distributions. We also anticipate that specific graphical displays in fields such as computer science, genetics, and social networks can be made more effective by explicitly displaying reference distributions alongside data displays, or by constructing plots of residuals or latent-data plots that would reveal structure over and above what is expected in fitted models.

On a more specific point, I thank Buja for pointing out the error in the claim that the Kolmogorov-Smirnoff test could be directly used to detect the misfit in our mixture model example. As he points out, performing this model check requires a more elaborate treatment of the replication distribution. Another possibility would be to use a directional discrepancy, $T(y, y^{\text{rep}}) = \sup_x (F_y(x) - F_{y^{\text{rep}}}(x))$, rather than the Kolmogorov-Smirnoff distance, $T^{\text{K-S}}(y, y^{\text{rep}}) = \sup_x |F_y(x) - F_{y^{\text{rep}}}(x)|$. Unlike $T^{\text{K-S}}$, the directional discrepancy T is antisymmetric in y and y^{rep} and has the property that, if the model is true, its distribution is symmetric about zero. However, for this example the directional discrepancy will not actually detect the model misfit, because it could be equally likely to be positive or negative here. So Buja is correct that more effort would be needed to numerically summarize this model misfit that is visually so clear. The suggestions in his discussion provide some interesting directions for general comparisons of distributions.

2. BAYESIAN DATA ANALYSIS AND GENERATIVE MODELS

Buja questions why I focus on Bayesian posterior predictive distributions to the near exclusion of other approaches for creating reference distributions such as permutation tests, bootstraps, and cross-validation. The immediate motivation is probably from seeing so many Bayesian analyses with the following pattern:

1. "Exploratory data analysis": simple displays of raw data; for example, histograms or scatterplot matrices.
2. Construction of a series of more complicated models.
3. Extensive discussion, often including graphical diagnostics, of the convergence of iterative simulations.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms

4. Presentation of parameter estimates and uncertainties (typically in simple tabular form), to conclude the analysis.

These analyses rarely feature model checking or model-based graphical displays. Sometimes there is model comparison or model averaging, featuring numerical measures such as BIC or DIC, but rarely a graphical check showing the implications of the entire fitted model.

In some sense, the rarity of model checks can be understood on sociological grounds: if the final model does *not* fit the data, researchers have an obligation to improve the model until it fits, at which point a diagnostic check would be unnecessary. However, we suspect that the real issue is that models are checked little if at all.

I believe that one reason for Bayesian models not being checked is the perception that model checking is not necessary or even appropriate in Bayesian inference—an attitude we associate with a superficial reading of Savage, Lindley, and other subjective Bayesians. In our recent research we have tried to provide a theoretical foundation and examples of Bayesian predictive model checking (see Gelman et al. 2003, chap. 6).

In short, this article focused on Bayesian model checking partly because of our own positive experiences with Bayesian inference, but also because it seemed to us that Bayesians had the most to gain from model checking. Non-Bayesians seem more aware of the potential problems of using wrong models, and we wanted to show Bayesians how effective exploratory data analysis can be used, if taken seriously as part of the iterative model-checking process.

More generally, our approach requires a *generative model*—that is, a probability model that has the potential to generate observed data along with hypothetical replications. As noted by Buja, posterior inference has the additional advantage of automatically generating a set of generative models, thus separating inferential from predictive uncertainty—but for most applications, we have found the predictive uncertainty to be the most important, which is why we suspect these approaches could be nearly as successful with maximum likelihood or bootstrap inference as with full Bayes.

However, we would anticipate more difficulty applying EDA approach to statistical methods that are not fully model-based. Here we are thinking of methods such as quasi-likelihood, generalized estimating equations, and probability-weighted estimates for censored data, which produce parameter estimates without explicitly modeling the data-generation process. (For example, marginal models in biostatistics for longitudinal data estimate unit-level regression coefficients without fully specifying a model for the observations at each time point.) For these methods, it is not so easy to generate hypothetical replications, or random imputations of missing or latent data, and so plots of type 1, 2, and 4 (see the beginning of this rejoinder) cannot be routinely implemented. This is one reason we prefer to use generative models, whether or not their parameters are estimated Bayesianly.

This content downloaded from 131.123.92.13 on Tue, 17 Oct 2017 05:21:04 UTC
All use subject to http://about.jstor.org/terms