A Bayesian formulation of exploratory data analysis and goodness-of-fit testing*

Andrew Gelman†

January 3, 2003

Abstract

Exploratory data analysis (EDA) and Bayesian inference (or, more generally, complex statistical modeling)—which are generally considered as unrelated statistical paradigms—can be particularly effective in combination. In this paper, we present a Bayesian framework for EDA based on posterior predictive checks. We explain how posterior predictive simulations can be used to create reference distributions for EDA graphs, and how this approach resolves some theoretical problems in Bayesian data analysis. We show how the generalization of Bayesian inference to include replicated data y^{rep} and replicated parameters θ^{rep} follows a long tradition of generalizations in Bayesian theory.

On the theoretical level, we present a predictive Bayesian formulation of goodness-of-fit testing, distinguishing between p-values (posterior probabilities that specified antisymmetric discrepancy measures will exceed 0) and u-values (data summaries with uniform sampling distributions). We explain that p-values, unlike u-values, are Bayesian probability statements in that they condition on observed data.

Having reviewed the general theoretical framework, we discuss the implications for statistical graphics and exploratory data analysis, with the goal being to unify exploratory data analysis with more formal statistical methods based on probability models. We interpret various graphical displays as posterior predictive checks and discuss how Bayesian inference can be used to determine reference distributions.

The goal of this work is not to downgrade descriptive statistics, or to suggest they be replaced by Bayesian modeling, but rather to suggest how exploratory data analysis fits into the probability-modeling paradigm.

We conclude with a discussion of the implications for practical Bayesian inference. In particular, we anticipate that Bayesian software can be generalized to draw simulations of replicated data and parameters from their posterior predictive distribution, and these can in turn be used to calibrate EDA graphs.

Keywords: bootstrap; Fisher's exact test; graphics; mixture model; model checking; multiple imputation; prior predictive check; posterior predictive check; p-value; u-value

*To appear in the *International Statistical Review*, and based on a paper presented at the Seventh Valencia Meeting on Bayesian Statistics. We thank Elja Arjas, Johannes Berkhof, Xiao-Li Meng, Michael Newton, Dalene Stangl, Iven Van Mechelen for helpful discussions, the Research Council of Katholieke Universiteit Leuven for Fellowship F/96/9 and Grant OT/96/10, and the U.S. National Science Foundation for grants SBR-9708424, SES-9987748, SES-9987748, and SES-0084368.

†Department of Statistics, Columbia University, New York, NY 10027, U.S.A., gelman@stat.columbia.edu, http://www.stat.columbia.edu/~gelman/.

1 EDA and Bayesian inference can cooperate

1.1 Background

Two areas in which statistics has made great advances in the last few decades are exploratory data analysis and complex probability modeling. One one side, exploratory methods have broadened the scope of statistics to include 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 et al., 1983, Cleveland, 1985, 1993, Tufte, 1983, 1990, Buja, Cook, and Swayne, 1996, Wainer, 1997, among others). At the same time, Bayesian methods have been developed to fit data much more realistically using hierarchical models with large numbers of parameters to model heterogeneity, interactions, and nonlinearity; see, for example, Gelman et al. (1995), Carlin and Louis (1996), and Denison et al. (2002) for recent reviews.

Interestingly, both these approaches have been fueled by computational improvements: for exploratory data analysis and data visualization, higher-resolution graphics and more sophisticated interactive user interfaces; for modeling, faster computers that allow routine use of iterative methods such as Markov chain simulation algorithms for fitting models with no closed-form expressions for estimates, uncertainties, and posterior distributions.

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 presented as model-free. 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.

We argue in this paper 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 set up a theoretical formulation in which graphical displays can be viewed as model checks, so that new models and new graphical methods go hand in hand.

On a practical level, we suggest to modelers that they check fit by comparing to replications of potential future data from the estimated model. Conversely, we suggest to exploratory data analysts that they proceed iteratively, using graphs at the initial phases of an analysis and also later on, to find patterns that represent deviations from the current state-of-the-art model.

1.2 EDA is based on models

Exploratory data analysis is typically presented as model-free. However, models are there as a baseline. For example, Tukey (1972) presents two-way fit plotting, in which an additive model is explicit; and hanging rootograms, which can be interpreted in terms of a Poisson rate for counts (although this model is not ever stated in the paper). In Tukey's words, exploratory plots are "graphs intended to let us see what may be happening over and above what we have already described."

Our proposal is to use complex models and Bayesian inference to advance "what we have already described" so that exploratory analysis becomes more powerful. Or, conversely, we seek to use the methods of exploratory analysis to check complex models and lead to ideas for their improvement.

1.3 Statistical graphics as model checking

We view *model checking* as the comparison of data to replicated data under the model. This includes "exploratory data analysis" and "confirmatory data analysis" as special cases: EDA is the graphical comparison and CDA is the p-value, but they are based on the same hypothesis test.

The 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 fitted model departs from the data. In a Bayesian model-building framework, EDA and CDA can both be applied at various stages in the analysis, including at a final stage—after any model selection or model averaging has been performed, posterior simulations can be computed from the final model.

2 Mathematical formulation

2.1 Bayesian inference: a history of generalizations

It is an important tradition in Bayesian statistics to formalize potentially vague ideas, starting with the axiomatic treatment of prior information and decision making from the 1920s through the 1950s. For a more recent example, consider hierarchical modeling. In the 1960s and 1970s, it was recognized that Bayesian inference for a sequence of parameters could have better statistical properties if data-dependent prior distributions were allowed. This developed into the "empirical Bayes" approach. But then, through the work of Hill (1965), Tiao and Tan (1965), Lindley and Smith (1972), Rubin (1981), and others, the hierarchical Bayes approach was developed, obtaining the benefits of data-based prior distributions in a fully Bayesian mathematical framework. Other places where vague statistical ideas have been formalized are in modeling missing data (Rubin, 1976) and model averaging (Draper, 1995, Raftery, 1995).

All of these ideas have the form of mathematical generalizations. Start with the likelihood, $p(y|\theta)$. Bayesian inference generalizes to include a prior distribution, $p(\theta)$. Over the years, many statisticians have objected to the claim that they need a prior distribution—but the success of Bayesian methods suggests that the gains (in ability to flexibly restrict inferences and to perform exact decision analyses) outweigh the costs that arise from having to defend "subjective" inferences. In fact, classical inferences can often be interpreted as Bayesian under particular prior specifications and loss functions, and so the Bayesian approach can be a tool to understand other statistical methods.

Similarly, hierarchical modeling elicited a lot of resistance in its time (see, for example, the discussion of Lindley and Smith, 1972), with a key point of contention being the legitimacy of combining information from different sources in a single model, as in a meta-analysis. There was also some free-floating skepticism about the additional assumptions inherent in an empirical Bayes analysis or hyperprior distribution[1]. Eventually, however, the intermediate formalism of "empirical Bayes," with its awkward data-dependent prior distributions, was replaced by the richer full-Bayes hierarchical structure. It became clear that the hierarchical analysis is a generalization that includes simpler models as special cases, and this allows us to answer various objections at a mathematical level. For example, if a hierarchical model combines highly dissimilar data sources—and these dissimilarities are not corrected for in the model—then the hierarchical variance parameter will be estimated to be a very large value, and the inferences will display essentially no shrinkage.

The next generalization, modeling missing data or, more generally, the process of data collection, generalizes the likelihood from $p(y|\theta)$ to $p(y, I|\theta, \phi)$, where I represents the information of which data points are actually observed, and ϕ are parameters describing the design of the data-collection and recording process (Rubin, 1976). Including the data-structure I in the model allows us to easily model rounded, censored, and truncated data and, as with the previous generalizations, gives insights into the previously-standard methods. In the more general framework, a model is "ignorable" if $p(\theta|y) = p(\theta|I, y)$; that is, if the data structure can be ignored. Understanding ignorability helps us in setting up non-ignorable models (as with dropouts in clinical trials) and in adding covariates to a model so that ignorability can be a reasonable assumption. Also as with the previous generalizations, these concepts predated the mathematical formalism, but the formalism made it easier to apply them in new and more complicated settings.

The expansion of the Bayesian formalism into $p(y, I|\theta, \phi)$ to include the data-generation process using $p(y, I)$ also resolves some theoretical and practical connections to classical methods (see Gelman et al., 1995, chapter 7). For example, randomized data collection is hard to justify under the usual Bayesian framework, but, in the context of defining a data-collection scheme, randomization

[1] I recall seeing a graduate student presentation a few years ago of a hierarchical regression model that had random effects for the 50 U.S. states. A statistician objected that the 50 states are fixed, and so it does not make sense for them to be random effects, in the sense of their being a larger population from which they are a sample. This is an interesting point but not relevant to the hierarchical model *per se*. One could similarly object to a non-hierarchical regression model of data from 50 states, since once again there is an error distribution. In either case, the model must be interpreted with care—but that there are only 50 states is not a good reason to set the state-level variance parameter to zero or infinity, as would be implied by classical nonhierarchical models.

is in fact the only way to select a sample without reference to covariates. Similarly, the idea of ignorability corresponds to the classical principle of including in the analysis all information used in the design, which in turn suggests particular Bayesian models. And the traditional Bayesian claim about the irrelevance of data-based stopping rules (see, for example, Berger, 1985) is modified by an understanding that a time variable must be included to have an ignorable model in this scenario.

A very active area of current statistical research is model averaging, generalizing the space of parameters one step further to allow for different choices of models or (in our preferred version) a continuous space spanned by models which had previously been fitted individually. Much progress seems to have been sparked by various formalizations of model combination, which take us beyond the previous vague ideas that no model is perfect and that it should be desirable to combine inferences from several models. Mathematical gaps typically correspond to areas of potential statistical improvement, and one area for improvement here can be seen from the difficulties of computing Bayes factors for models of different dimensionality (see, for example, Raftery, 1995, Spiegelhalter et al., 2002, and Denison et al., 2002). The problem here is not with the model combination but rather with the use of flat, or nearly-flat, prior distributions on the component models. We suspect that model averaging would be much more effective if the models being averaged were hierarchical.

A classic gambit of Bayesians is to claim that other statistical methods are Bayesian too, merely with unstated (and probably incoherent and unrealistic) prior distributions. This line of argument is not just rhetorical—it has motivated interesting research (e.g., Box and Tiao, 1973, Wahba, 1978, Clyde and George, 2000)—and we use it in developing Bayesian analogues to classical goodness-of-fit tests.

2.2 Model checking

Gelman, Meng, and Stern (1996) made a case that model checking warrants a further generalization of the Bayesian paradigm, and we continue that argument here. The basic idea is to expand from $p(y|\theta)p(\theta)$ to $p(y|\theta)p(\theta)p(y^{\text{rep}}|\theta)$, where y^{rep} is a replicated data set of the same size and shape as the observed data y. All model checking (both "exploratory" and "confirmatory") can then be interpreted as comparisons between y and y^{rep}.

As with other generalizations of Bayesian formalism, the y^{rep} notation makes an existing procedure explicit. The need to precisely define a replication distribution, $p(y^{\text{rep}}|\theta)$—like the earlier need to define a prior distribution—implies additional effort which is intended to pay off in the form of more precise inferences.

More generally, we consider replications of the parameters too, hence the full Bayesian model, as we see it, is $p(y, y^{\text{rep}}, \theta, \theta^{\text{rep}})$, and all posterior calculations (including model checks) use the distribution, $p(y^{\mathrm{rep}}, \theta, \theta^{\mathrm{rep}}|y)$. Of course, the complexities of notation described in the previous section (for hierarchical parameter structures, missing data, and model averaging) can be folded in here as appropriate (see Gelman et al., 2002).

Our most general form of test involves antisymmetric "discrepancy" functions of data and replications of the form $D(y, y^{\mathrm{rep}}, \theta, \theta^{\mathrm{rep}})$. These are antisymmetric in the sense that if (y, θ) are exchanged with $(y^{\mathrm{rep}}, \theta^{\mathrm{rep}})$, then D must change sign. The most common special case arises from test variables of the form $T(y, \theta)$, in which case the antisymmetric discrepancy can be defined as $D(y, y^{\mathrm{rep}}, \theta, \theta^{\mathrm{rep}}) = T(y^{\mathrm{rep}}, \theta^{\mathrm{rep}}) - T(y, \theta)$. The advantage of working with antisymmetric discrepancies rather than test statistics is that the discrepancies are always compared to zero, which should allow visual model checks to be clearer (Berkhof, Van Mechelen, and Gelman, 2002).

2.3 U-values and P-values

In light of the recent confusion in the statistical literature about Bayesian p-values, some definitions may be in order. We define a Bayesian p-value as a posterior probability under a particular modeling assumption. We move from a classical definition,

$$\text{p-value}(y|\theta) = \Pr(T(y^{\mathrm{rep}}) > T(y) \,|\, y, \theta),$$

to the Bayesian version, averaging over the posterior distribution of θ:

$$\text{p-value}(y) = \Pr(T(y^{\mathrm{rep}}) > T(y) \,|\, y).$$

More generally, we can consider any antisymmetric discrepancy function of the form $D(y, y^{\mathrm{rep}}, \theta, \theta^{\mathrm{rep}})$, and then,

$$\text{p-value}(y) = \Pr(D(y, y^{\mathrm{rep}}, \theta, \theta^{\mathrm{rep}}) > 0 \,|\, y).$$

This can be generalized further by considering missing and latent data (and thus involving the data-collection indicator I) or model averaging, but the basic idea still holds: a p-value is a Bayesian posterior probability that a certain antisymmetric function exceeds zero.

In the special cases in which pivotal test statistics exist, the classical p-value also has the property of having a Uniform$[0, 1]$ distribution, considering y as a random variable under the model. We can generalize this to define the *u-value* as any function of the data y that has a uniform sampling distribution. Although the u-value can be defined Bayesianly, by averaging over the distribution of θ, it is fundamentally *not* a Bayesian quantity, in that it cannot be interpreted as a posterior probability statement about an underlying truth. (In contrast, the p-value is a statement, conditional on the model, about what might be expected in future replications.)

The p-value is to the u-value as the posterior interval is to the confidence interval. Just as posterior intervals are not, in general, classical confidence intervals, Bayesian p-values are not generally

u-values. We prefer to work with posterior intervals and p-values (rather than confidence intervals and u-values) because of their direct interpretations in terms of posterior probabilities.

2.4 Prior predictive checks are posterior predictive checks too

Prior predictive checks (Box, 1980) are sometimes taken as a "purer" alternative to posterior predictive model checking. In a prior predictive check, the replicated data y^{rep} are drawn from their prior distribution, $p(y^{\text{rep}}) = \int p(y^{\text{rep}}|\theta)p(\theta)d\theta$. On a practical level, prior predictive checks have the problem that they are attempting to test the entire model—including those parameter values θ that are essentially ruled out by the data. In contrast, posterior predictive checks are a natural generalization of the standard classical approach of plugging a point estimate for θ into a model check.

If we are careful in how we define replications, however, we can think of prior predictive checks as a subset of posterior predictive checks. In a posterior predictive check, we are generalizing to future data generated from the same parameter θ, whereas in a prior predictive check, θ^{rep} is redrawn from the model. And this is simply posterior inference about a different scenario. In a setting where the prior check "rejects" the model but the posterior check does not, this just means that the data are consistent with some some of the immediate implications of the model but not some of its more distant implications. We discuss this point further in the next section.

3 Theoretical examples

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).

Sections 3.1 and 3.2 illustrate with two examples where the inherent difficulties of a model are revealed by comparing data to predictive simulations. In both these cases, the problems of the models are well known in the statistical literature while at the same time subtle enough to frequently trap practitioners into mistakes. Then, in Section 3.3 we consider how clear definitions of replication distributions allow us to resolve seeming ambiguities in the classical problem of model checking with contingency tables.

3.1 Finite mixture models

Our first example is the fitting of a mixture model with unconstrained variances to continuous univariate data. A relatively simple form of this model has two equal components, with mixture density

$$p(y_i|\mu_1, \mu_2, \sigma_1, \sigma_2) = 0.5\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{1}{2\sigma_1^2}(y_i-\mu_1)^2} + 0.5\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{1}{2\sigma_2^2}(y_i-\mu_2)^2}. \tag{1}$$

When fit to data $y_i, i = 1, \ldots, n$, using maximum likelihood, a problem arises: 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 even if the model is true!

Unfortunately, this problem is *not* immediately solved through Bayesian inference. For example, if we assign (improper) Uniform$(-\infty, \infty)$ prior densities for μ_1, μ_2, and Uniform$(0, \infty)$ prior densities for σ_1, σ_2, then the posterior modes will still be at the points where one or another of the σ's approach 0, and these modes in fact contain infinite posterior mass—the Bayesian averaging over uncertainty does not save us.

But now consider attacking this problem using the Bayesian approach that includes inference about y^{rep} as well as θ. In practice, this means summarizing the posterior distribution of θ (possibly using iterative simulation) and then, for each simulation of θ, simulating a new vector y^{rep} of independent draws from the mixture distribution. There are two likely possibilities:

1. At least one of the modes (with their 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. 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, since the computational procedure is in effect fitting a truncated model that fits the data well.

In either case, posterior predictive checking has worked in the sense of "limiting the liability" caused by fitting an inappropriate model. In contrast, a key problem with Bayesian inference—if model checking is not allowed—is that if an inappropriate model is fit to data, it is possible to end up with highly precise, but wrong, inferences.

3.2 Random effects models and overfitting

Can we tell that a model is "overfitting" a dataset? Sometimes, and it depends on the context. We explore with the simple hierarchical normal model for the one-way data structure. Rubin (1981) and Gelman et al. (1995, chapter 5) describe data from randomized experiments on test-coaching programs in 8 schools: for each school $j = 1, \dots, 8$, there is an unbiased estimate y_j of the effectiveness of the program in the school, along with an (essentially) known standard error, σ_j. The goal is to estimate the true treatment effects, θ_j, in the 8 schools.

We all know that the simplest analysis of these data—analyzing each of the schools independently and coming up with estimates $\hat{\theta}_j = y_j$—is inappropriate. Shrinkage is better (see, for example, Efron and Morris, 1975). But what if you had never heard of Stein (1955) or the rest of the literature, and you just naively fit the simple model? Could posterior predictive checking tell you about your mistake—in this case, a mistake of overfitting, using 8 independent parameters when more structure is needed?

In a word, yes. The problem with the unshrunk estimator, as with overfitting in general, is that the parameters capture too much of the variability in the data. This can be simply captured using the sample variance of the data as a test statistic.

3.3 Contingency tables and "exact" tests

What should be conditioned on when testing hypotheses in a contingency table? Is "Fisher's exact test" actually exact? To answer this seemingly unanswerable question, we must think carefully about the replication distribution.

So-called exact permutation tests are appropriate if considering experiments with fixed margins. This is unusual in practice. It is far more common to have one or zero margins fixed:

1. For example, consider an experiment with 4 treatment groups and 3 outcomes. You might do the experiment with a fixed number of persons in each group, but the outcomes are random variables. In the replicated data, too, the outcomes should be random.

2. For a case-control study, the outcome margins might be fixed but the treatment margins are random.

3. In observational studies, it is common for both margins to be random: data are collected on n persons, and then the treatment and outcome states of each are recorded. Once again, a

replicated data set should follow this design.

4. Fisher's tea-tasting experiment is one of the very few in which both margins are fixed by design (the lady is given 4 cups of each kind of tea, and she is told ahead of time that there are 4 of each, so that her guesses will be balanced also).

What is the difficulty, then? For the appropriate test, conditioning on only one of the margins means that the other is random, and so the reference distribution, and thus the test itself, depends on the unknown parameter θ representing those marginal frequencies. But this is not a problem in a Bayesian context—we just average over the posterior distribution of θ, and then over the replication distribution of y^{rep}. People have wasted a lot of time trying to figure out how to sample from the distribution of the discrete counts conditional on both margins, but that's just an inappropriate calculation in almost all cases.

To put it another way, Fisher's exact test gives a u-value that is not in general a p-value (except in the highly unusual scenario of an experiment with both margins fixed by design). From a Bayesian perspective, the permutation test can be justified as a convenient approximate inference, by analogy to the way in which maximum likelihood estimation is often more convenient than working with a full posterior distribution. But in problems where calculating the permutation distribution is difficult, the convenience is lost, and then we strongly recommend going back to first principles, defining a replication distribution based on the actual data collection process, and computing an actual p-value. Once again, this follows the analogy of Bayesian inference: when likelihood analysis becomes complicated (for example, multimodal or constrained likelihoods), it is ultimately simpler to set up a prior distribution and do the full Bayesian analysis.

Interestingly, by including the distribution for y^{rep} in our Bayesian inference, we are following the classical statistical principle of accounting for the data collection in the analysis. At the same time, this analysis *does* follow the likelihood principle. The likelihood, $p(y|\theta)$, and the prior distribution, $p(\theta)$, do not depend on the data-collection process—but the predictive distribution for the replicated data, $p(y^{\text{rep}}|\theta)$, is affected by the design. The likelihood principle (see, e.g., Berger, 1985) does not apply to y^{rep}, which is important since these are the future data that we would like our model to predict.

4 Towards a theory of exploratory data analysis

4.1 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. At the other extreme, 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 here is thus to columns of equal width, which is not particularly interesting from a substantive perspective in that example.

Some of the most useful and interesting systematic research on statistical graphics (for example, Ehrenberg, 1975, Tukey, 1977, Tufte, 1983, and Cleveland, 1985) has paralleled the research methods of psychology, in particular:

1. Assessing the information content in a table or graph—this is similar to work in mathematical psychological models of information and perception;
2. Comparing the understandability of the same data graphed different ways, following the principles and methods of experimental psychology; and
3. Introspection, which as in psychological research is still an extremely powerful (and convenient) tool in working out examples and trying to figure out why some graphical displays tell the story better than others.

In parallel with the attempts at synthesis have come developments of new graphical methods (for example, Chambers et al., 1983, Tufte, 1990, and Cleveland, 1993, just to restrict ourselves to static graphical displays).

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. Once we systematically think of graphs as model checking, we can think of ways that a graphical display can take advantages of symmetries in the reference distribution of $T(y^{\rm rep}, \theta)$. Or, conversely, how certain graphical displays can be misleading because they implicitly *assume* symmetries that are inappropriate to the model being considered.

4.2 Adapting graphical forms to the structures of test statistics

1. The most basic exploratory graphic is simply a display of an entire data set (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 need, as comparisons, displays of $T(y^{\rm rep})$ corresponding to several draws from the reference distribution.

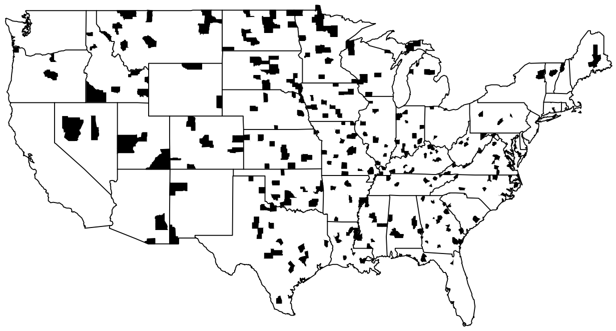

Figure 1: The 10% of counties of the United States with the highest age-standardized death rates for cancer of kidney/ureter for U.S. white males, 1980–1989. The most notable pattern in this map (that most of the shaded counties are in the center-west of the country) can in fact be explained as an artifact caused by varying sample sizes. From Gelman and Nolan (2002).

Figure 1 shows the perils of attempting to interpret data *without* comparing to a reference distribution—the apparent patterns in this and similar 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 values (see Gelman and Price, 1999, for more on the cancer rate example and Louis, 1984, for discussion of the general issue of summarizing the uncertainty in an ensemble of parameters).

Our point here is not that this problem is solved by any simple Bayesian method—but rather that the act of looking at a map for patterns is itself implicitly a comparison to a patternless reference model. Because of the spatial variation of populations of counties, the comparison to a patternless map is essentially meaningless.

2. 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) discusses 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) presents a striking example. In many applications involving structured data, we have found that a single replicated dataset would almost be enough to give a convincing picture of how the model is not fitting the data.

3. At the opposite extreme, if we have a scalar test summary, we can overlay it on a histogram of its simulated reference distribution. A two-dimensional summary can similarly be shown in comparison to a scatterplot.

4. A multidimensional summary, $T(y) = (T_1(y), \dots, T_k(y))$, can be shown as a scatterplot of $T_k(y)$ vs. k, in comparison with several scatterplots of $T_k(y^{\text{rep}})$ vs. k. But this comparison can

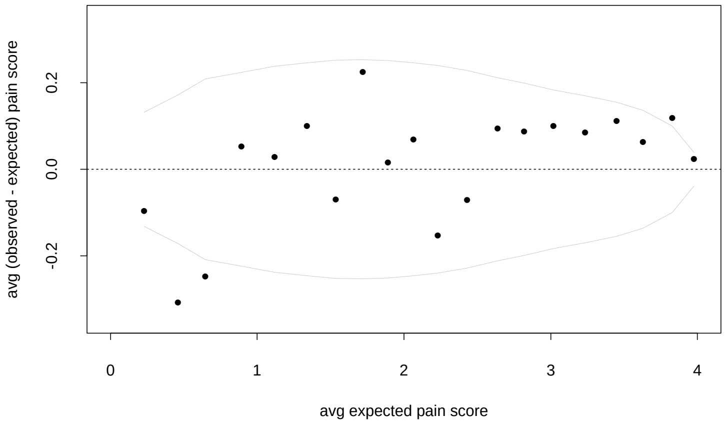

Figure 2: Average residuals vs. expected values, with discrete responses divided into 20 equally-sized bins defined by ranges of expected values, for a nonlinear model of an ordered categorical data set. The prediction errors are relatively small but with a consistent pattern that low predictions are too low and high predictions are too high. Dotted lines show 95% bounds under the model. Adapted from Gelman and Bois (1997).

be displayed much more compactly using line plots: a single graph can show the line of $T_k(y)$ vs. k in bold, overlaying several lines of $T_k(y^{\text{rep}})$ vs. k, each corresponding to a different draw from the reference distribution.

5. The above plots can be usefully simplified if the reference distribution has certain invariance properties. For example, consider a binned residual plot of $\bar{r}_k$ vs. $\bar{u}_k$, for bins $k = 1, \ldots, K$, as in Figure 2. We also removed the lines connecting the dots for the data residuals, since there is no longer a background of replication lines. Instead, comparison is to the implicit independence distribution.

   Under the reference distribution, the residuals are independent and, if enough are in each bin, the mean residuals $\bar{r}_k$ are approximately normally distributed. We can then display the reference distribution as 95% error bounds, as in Figure 2. More discussion of Bayesian binned residual plots appears in Gelman et al. (2000).

6. Hierarchical structure in a model 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. Figure 3 shows an example of poor fit (clearly revealed by a single simulation draw of the parameter vectors). The model was altered, and the new check appears in Figure

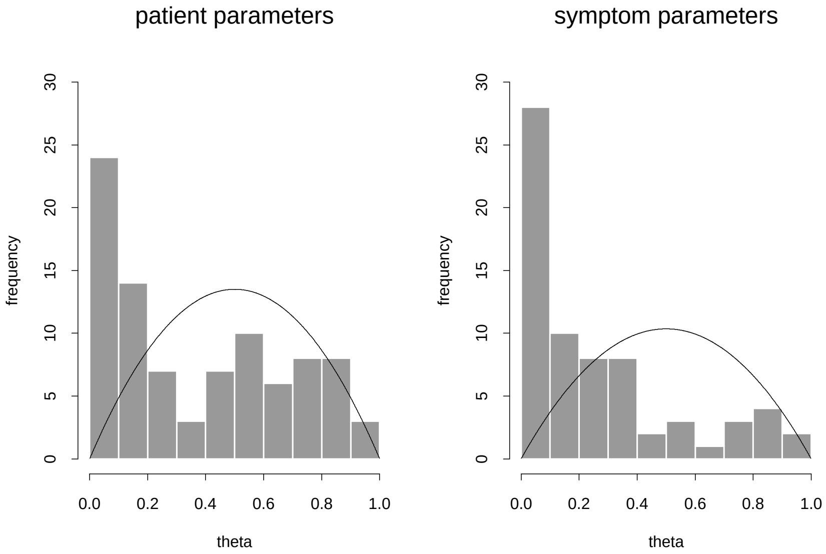

Figure 3: Histograms of (a) 90 patient parameters and (b) 69 symptom parameters, from a single draw from the posterior distribution of a psychometric model. These histograms of posterior estimates contradict the assumed Beta$(\theta|2,2)$ prior densities (plotted on top of the histograms) for each batch of parameters, and motivated us to switch to mixture prior distributions. This implicit comparison to the values of θ_j under the prior distribution can be viewed as a posterior predictive check in which the replicated data include 30 new patients and 23 new symptoms. From Meulders et al. (1998).

4. This could be considered a "manual Gibbs sampler," in which aspects of the model are altered to bring them in line with the data.

7. In other cases, a reference distribution is implied, not from symmetries in the model or test statistic, but from prior information that has not been included in the model. Gelman et al. (2002) show examples in which simulated completed data—from fitted models—are displayed. Unexpected patterns can be attributed to misfit of the model or to unexpected features in reality revealed by the model fit. It can be difficult to visually judge the statistical significance of some of the fluctuations, which is where simulations from the posterior predictive distribution can be used as comparisons.

5 Conclusions, or, isn't this all obvious?

Further research is needed to connect the principles of graphical design (as in Tufte, 1983, 1990, Chambers et al., 1983, and Cleveland, 1985, 1993) to the formal ideas of comparing data and antisymmetric discrepancies to reference distributions. Already we have reinterpreted some well-known ideas, such as binned residual plots (now justified because they can be implicitly compared to

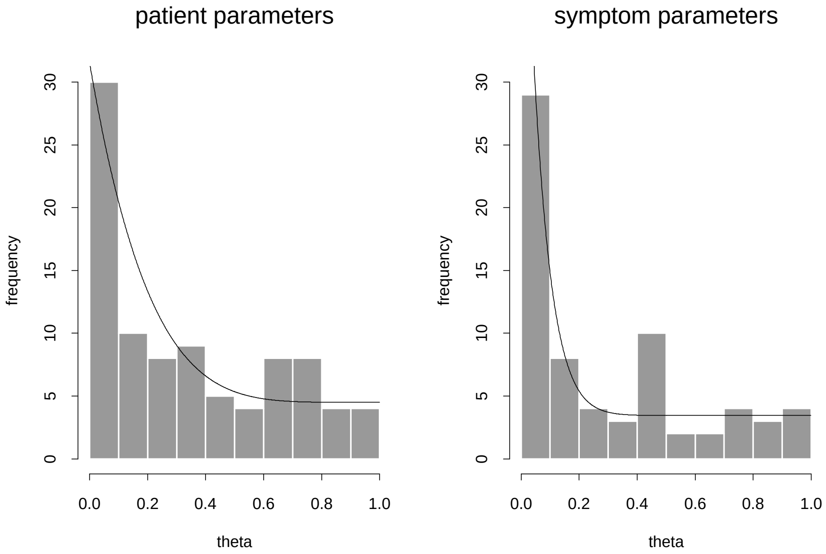

Figure 4: Histograms of (a) 90 patient parameters and (b) 69 symptom parameters, as estimated from the expanded psychometric model. The mixture prior densities (plotted on top of the histograms) are not perfect, but they approximate the corresponding histograms much better than the Beta($\theta|2, 2$) densities in Figure 3. From Meulders et al. (1998).

their symmetric and independent reference distributions) and come up with new ideas appropriate to complex models (for example, the idea of plotting a single simulation draw of a batch of hierarchical parameters, rather than displaying the marginal posterior distribution of each). The former allows direct comparison to a replication distribution, as in Figures 3 and 4.

We expect there is room for improvement and for future statistical packages to have automatic features for simulating replication distributions and performing model checks. We can anticipate three challenges:

1. The replication distribution. This is analogous to the problem of specifying the prior distribution in a Bayesian analysis. It can never be automatic, but standard options will be possible. For example, in a language such as BUGS (Spiegelhalter et al., 1994, 2002), 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 might require more effort in setting up the model, which is as it should be. For example, if a sample size parameter n is to be resampled, it will need to have a distribution specified, and this distribution could depend on the data, as would be appropriate if analyzing data from a sequential design.

2. The test variables. Presumably, one can start by picking a bunch of these. It would be natural to choose various summaries of any batches of random effects, for example.

3. Graphical display. These would have to be adapted to the dimensionality and structure of the test variables and their replication distributions, as discussed in Section 4.2.

In the Bayesian analyses we have seen, exploratory data analysis is often performed at the beginning, and there is sporadic model checking later. We hope that through the explicit inclusion of the replications as part of the model, more sophisticated model checks will become standard procedure. These will range from automatic quantitative comparisons (for example, checks for skewness and autocorrelation) to graphical explorations of the data, to discover subtle patterns of the data using complex Bayesian model fits as a guide.

We conclude with two examples from the very recent Bayesian literature that illustrate the graphical exploratory analysis in the context of complex models. Carlin and Banerjee (2003) developed a new multivariate spatio-temporal correlation model and fit it to data. They report a model-checking display: "Year-by-year boxplots of the posterior median frailties (not shown) indicate a slightly decreasing trend ... though this might be mostly an artifact of the paucity of observed survival times for these most recent cohorts." We suspect that displayed draws of the vectors of the posterior time series would allow direct comparison to the replication distribution. In another example, Newton (2002) fitted a complex model to genetic data. He uses this model not just to make inferences about parameters but to graphically reexpress the data to which the model was fit (in his Figure 6). The next logical step would be to simulate replicated data under the model and compare these to the observed data, using this visual display.

In both these cases, the key steps are the scientific model that was already fit to the data, and the Bayesian inference used to draw inferences in the highly structured parameter spaces. Given this scientific, statistical, and computational effort, it is a small step to graph the data and parameter inferences and compare them, explicitly or implicitly, to replicated data. This step is important if there is interest in applying these models to new data.

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*, second edition. New York: Springer-Verlag.

Berkhof, J., Van Mechelen, I., and Gelman, A. (2002). Posterior predictive checking using antisymmetric discrepancy functions. Technical report, Department of Psychology, Katholieke Universiteit Leuven, Belgium.

Box, G. E. P. (1980). Sampling and Bayes inference in scientific modelling and robustness. *Journal of the Royal Statistical Society A* **143**, 383–430.

Box, G. E. P., and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis.* New York: Wiley Classics.

Buja, A., Cook, D., and Swayne, D. (1996). Interactive high-dimensional data visualization. *Journal of Computational and Graphical Statistics* **5**, 78–99.

Buja, A., Cook, D., and Swayne, D. (1999). Inference for data visualization. Talk given at Joint Statistical Meetings.

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

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

Cleveland, W. S. (1993). *Envisioning Information.* Summit, N.J.: Hobart Press.

Carlin, B. P., and Banerjee, S. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data. In *Bayesian Statistics 7*, to appear.

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

Clyde, M., and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society B.*

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.

Draper, D. (1995). Assessment and propagation of model uncertainty (with discussion). *Journal of the Royal Statistical Society B* **57**, 45–97.

Efron, B., and Morris, C. (1975). Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* **70**, 311–319.

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.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian Data Analysis.* 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., 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.* Oxford University Press.

Gelman, A., Pasarica, C., and Dodhia, R. (2002). Let's practice what we preach: turning tables into graphs. *The American Statistician.*

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. (2002). Bayesian model checking for missing and latent data problems using posterior predictive simulations. Technical report, Department of Statistics, Columbia Univeristy.

Hill, B. M. (1965). Inference about variance components in the one-way model. *Journal of the American Statistical Association* **60**, 806–825.

Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society B* **34**, 1–41.

Louis, T. A. (1984). Estimating a population of parameter values using Bayes and empirical Bayes methods. *Journal of the American Statistical Association* **78**, 393–398.

Meulders, M., Gelman, A., Van Mechelen, I., and De Boeck, P. (1998). Generalizing the probability matrix decomposition model: an example of Bayesian model checking and model expansion. In *Assumptions, Robustness, and Estimation Methods in Multivariate Modeling*, ed. J. Hox.

Newton, M. A. (2002). Discovering combinations of genomic alterations associated with cancer. *Journal of the American Statistical Association* **97**, 931–942.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63**, 581–592.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics* **6**, 377–401.

Spiegelhalter, D. J., Best, N., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, to appear.

Spiegelhalter, D., Thomas, A., Best, N., Gilks, W., and Lunn, D. (1994, 2002). BUGS: Bayesian inference using Gibbs sampling. MRC Biostatistics Unit, Cambridge, England. `www.mrc-bsu.cam.ac.uk/bugs/`

Stein, C. (1955). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley Symposium* **1**, ed. J. Neyman, 197–206. Berkeley:

University of California.

Tiao, G. C., and Tan, W. Y. (1965). Bayesian analysis of random-effect models in the analysis of variance. I: Posterior distribution of variance components. *Biometrika* **52**, 37–53.

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

Tufte, E. R. (1990). *Envisioning Information.* Cheshire, Conn.: Graphics Press.

Tukey, J. W. (1972). Some graphic and semigraphic displays. In *Statistical Papers in Honor of George W. Snedecor*, ed. T. A. Bancroft. Iowa State University Press.

Tukey, J. W. (1977). *Exploratory Data Analysis*, New York: Addison-Wesley.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society B* **40**, 364–372.

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