Discussion Article

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.

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

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.

- *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:

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

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

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.

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.