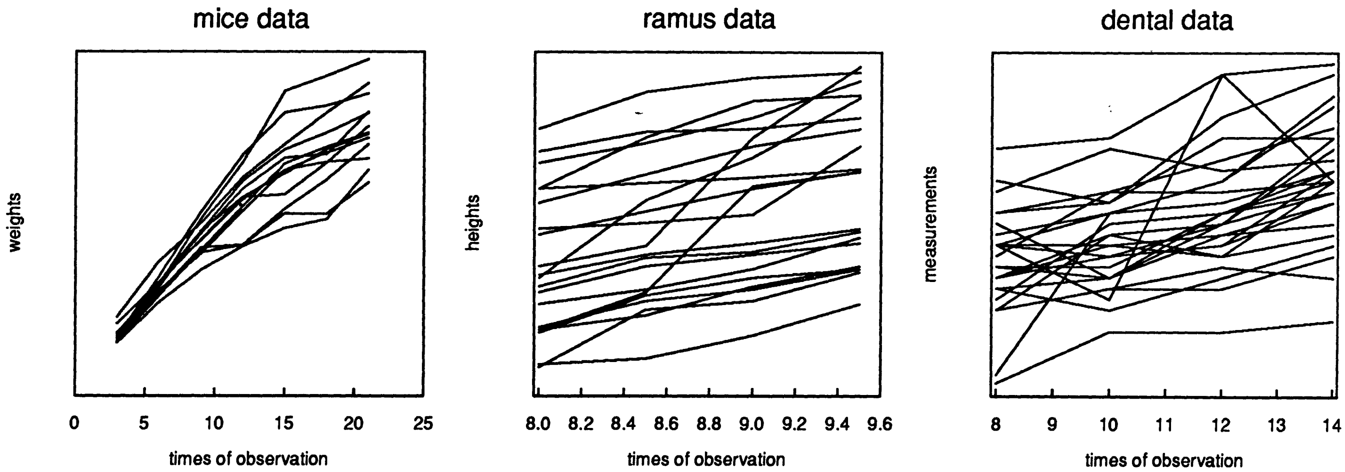

FIG. 2. *Plots of example data.*

where

$$g^* = \left[\frac{(n+1)(s-k)}{(n+1)(s-k)+2}\right]\left[\frac{n-k-2}{n}\right].$$

In conclusion, we agree with Professor Rao that his empirical Bayes predictor of future observations in growth curve models performs better than its least squares counterpart. We have also described several other empirical Bayes prediction methods. With the three example data sets, we have found our calibrated empirical Bayes predictor to yield smaller CVAE and to be more stable than its calibrated least squares counterpart.

ACKNOWLEDGMENTS

This work was supported by National Institutes of Health Grant GM29745. We also thank the Statistics Center, Massachusetts Institute of Technology, for the use of their computing resources.

ADDITIONAL REFERENCES

AMEMIYA, Y. (1985). What should be done when an estimated between-group covariance matrix is not nonnegative definite? *Amer. Statist.* **39** 112–117.

GOLDSTEIN, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika* **73** 43–56.

LAIRD, N., LANGE, N. and STRAM, D. (1987). Maximum likelihood computations with repeated measures: Application of the EM algorithm. *J. Amer. Statist. Assoc.* **82** 97–105.

LANGE, N. and LAIRD, N. M. (1986). Random-effects and growth-curve modeling for balanced and complete longitudinal data. Technical Report, Dept. Mathematics, Massachusetts Institute of Technology, Cambridge, Mass.

LONGFORD, N. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. Technical Report 87-77, Educational Testing Service, Princeton, N. J.

REINSEL, G. (1985). Mean squared error properties of empirical Bayes estimators in a multivariate random-effects generalized linear model. *J. Amer. Statist. Assoc.* **80** 642–650.

Comment: On Exchangeability Judgments in Predictive Modeling and the Role of Data in Statistical Research

David Draper

Professor Rao has shared with us some thought-provoking ideas on prediction in growth curve modeling. The paper has four basic attributes, two of which seem positive and two negative. On the positive side,

- the basic problem is predictive in nature, thereby emphasizing inference on observable quantities (future values of outcome variables of interest) rather than on unobservable quantities (parameters); and

David Draper is a member of the Statistical Research and Consulting Group, Department of Economics and Statistics, RAND Corporation, 1700 Main Street, Santa Monica, California 90406.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *Statistical Science.* www.jstor.org

- there is a strong emphasis on cross-validation to aid in the selection of what are essentially smoothing parameters in Rao's methods (his choices of degree of polynomial approximations to growth curves and number of relevant time periods over which the smoothing is to be done), and to supply uncertainty assessments.

On the other hand,

- the paper is speculative in nature, with many ideas, not all of them completely worked out. This would be fine for a philosophy of inference paper but is not so fine in a methodology article. The paper's general theory of Section 2 does not connect well with its methodology and applications, the cross-validation idea being the only one that carries over systematically from the second section to the other parts; and
- the paper features a curious use of data.

As an additional point of disagreement, Rao and I appear to have rather different views on how to approach data of the type he considers, particularly in the matter of forming judgments about what aspects of the past and future are conditionally exchangeable, in order that learning from past experience may occur. I will begin with some comments about the role of data in statistical research and will conclude with a discussion of the part played by exchangeability judgments in predictive modeling.

1. ON THE ROLE OF DATA IN STATISTICAL RESEARCH

In his comments on the otherwise excellent paper by Hastie and Tibshirani in this journal last year, Brillinger (1986) offered the following criticism in a section called "Some Quibbles":

> Two medical data sets are analyzed [in the Hastie and Tibshirani paper], but no inferences are made. Can the authors not set down some (biological) insight or understanding that has been gained from the analyses? Otherwise they might just as well have presented the results of simulations.

I want to expand on this point, which to me is more serious than a quibble, by using the present paper as one example of a fairly general feature of statistical writing.

How are data used in statistical research? There appear to be four main cases to distinguish.

(1) In the first case, a body of theory is developed or extended without any recourse ever being made to data, or an applied problem giving rise to data. There are many examples of this, a large number of them having appeared over the years in *The Annals of (Mathematical) Statistics*. Some of this work turns out ultimately to be useful, in the same way that mathematics for its own sake has repeatedly in the past been discovered fifty or a hundred years after the fact to be just what was needed to solve a real-world problem. But without the potentially firm footing of usefulness supplied by a serious applied problem, there can be no guarantees. (The two-by-two table with margins "applied versus theoretical" and "useful versus not useful" also has another pejorative cell—applied work that is not of much general usefulness because the solution was too narrowly tailored to the specific problem—but that's another story.) Rao himself draws a distinction between this first case and the other three below, as can be seen from this quote from his recent interview in *Statistical Science* (DeGroot, 1987):

> I always develop methodology from the data given to me for analysis rather than look at others' work and try to extend it in terms of mathematics.

Readers of this article who study Rao's data sets a bit will have to judge for themselves the extent to which his methods have arisen from his data on this occasion.

(2) In the second case, someone develops a methodological idea from first principles, and then looks for data to illustrate it. One or two data sets fitting the methodology's general template are pulled off the shelf, usually from other publications. In this case, the data are used principally to give examples of the calculations. People often say that papers of this sort have "real" data in them, as if this were a virtue. But since no serious attempt is made to look at the data or learn about the substantive issues, this is basically just window-dressing. For the purpose of illustrating the calculations, one might have done just as well or even better with contrived data that make the arithmetic easy, as in Searle (1971). Moreover, as Brillinger points out, a different kind of fake data—simulated according to a known mechanism—may also be more useful, for another reason: the behavior of the method with known inputs can then be studied. (This is one of the few really valuable uses of pure frequentist reasoning, in fact.) Rao's paper appears to fall into this category.

(3) The third case is one in which the writer tries to take the data seriously, but in a partial or complete substantive vacuum. Useful plots and numerical summaries are examined, and a serious effort is made to have the data suggest the appropriate methodology rather than the other way around, but key background information about the substantive issues is scarce or not well utilized. Much of the more thoughtful literature in exploratory data analysis (Velleman and Hoaglin, 1981; Hoaglin, Mosteller and Tukey, 1983, 1985) is of this form. Such work can be quite useful in serving as case studies of good data analytic practice for people looking at their own data who *do* know a lot about the

substantive issues, so this is better than (2), but it is still possible for fairly silly conclusions to result from insufficient grounding in substantive reality.

The ultimate example of this third case, and what is both good and bad about work in this category, is the book *Data* (Andrews and Herzberg, 1985), which consists of 71 data sets each graced with from 1 to 7 pages (median 1 page) of narrative describing the substantive issues. From the profession's point of view it is good to have sources like this, so that smart people can give us data analytic insights in new contexts, rather than just beating to death the stack loss and Minitab trees data (statistician heal thyself: I have been as guilty of this as anyone). But this is also potentially dangerous, for the same reasons pure exploratory data analysis (EDA) in the absence of substantive knowledge can be dangerous (a well known statistician once characterized a well known EDA expert to me as follows: "Give him 100 numbers out of your favorite random number generator without telling him about the lack of underlying structure, and he will find 10 interesting things about them that suggest structural insight, known in this case to be spurious. What, then, about the next 100 numbers that walk in his door?").

(4) In the fourth case, the statistician, usually working together with other investigators collaboratively, tries to learn a lot about the background issues and conduct a good preliminary data analysis, and then lets the methodology flow from these two. Examples of this category include Mosteller and Wallace's (1984) detailed, insightful investigation into the authorship dispute surrounding the *Federalist Papers*, and some articles in the applications section of the *Journal of the American Statistical Association* (DuMouchel and Harris, 1983; Weisberg, 1986; Lagakos, Wessen and Zelen, 1986; Reinsel and Tiao, 1987, for instance). As anyone who tries to document this category's scarcity will quickly find, however, there are precious few examples to mention, to the discredit of the statistics profession.

One reason for this is that the medium by which we convey most of our research findings to each other—journal articles—is not well suited to fully describing work of this type. It is difficult to fit all of the needed background information and details on why particular actions were taken in an applied context into a single journal article, and extended case studies like that of Mosteller and Wallace typically do not break up neatly into a *series* of such articles, either. In my view what we ought to be doing a lot more of as a profession is writing detailed case studies, book length (like Mosteller and Wallace) when the problem warrants it, of how we actually *did* things in applied contexts, warts and all. People could then read these narratives and make comparative judgments, after adjusting for relevant differences, about what aspects of past experience can be usefully borrowed in present and future situations.

The relevance of this discussion to the paper at hand is that it is difficult to know what to make of the paper's general conclusions drawn on the basis of its data sets, about which the author has shared so few insights with us. If Rao had just used the data to illustrate his prediction and cross-validation techniques numerically, fine; but he wants more—he wants to draw general conclusions about how to make predictions in growth curve models. For that purpose, multiple simulated data sets with known characteristics would have been more useful. Or, if he really did want to look carefully at these data sets, we might all have gained some worthwhile insights about mouse growth patterns or dental measurements, but he appears not to have tried very hard to find out what these data were trying to say, and (because the data sets were 6 to 18 years old and simply borrowed from other publications) he wasn't in a good position to draw on substantive knowledge.

It is a risky business to suggest that someone has not taken his data seriously without actually having watched him prepare his paper. But the methods in this article do not appear to have arisen out of a process that began with activities as basic as plotting individual growth curves (Figures 1 to 3) and examining them thoroughly: one of the paper's methods

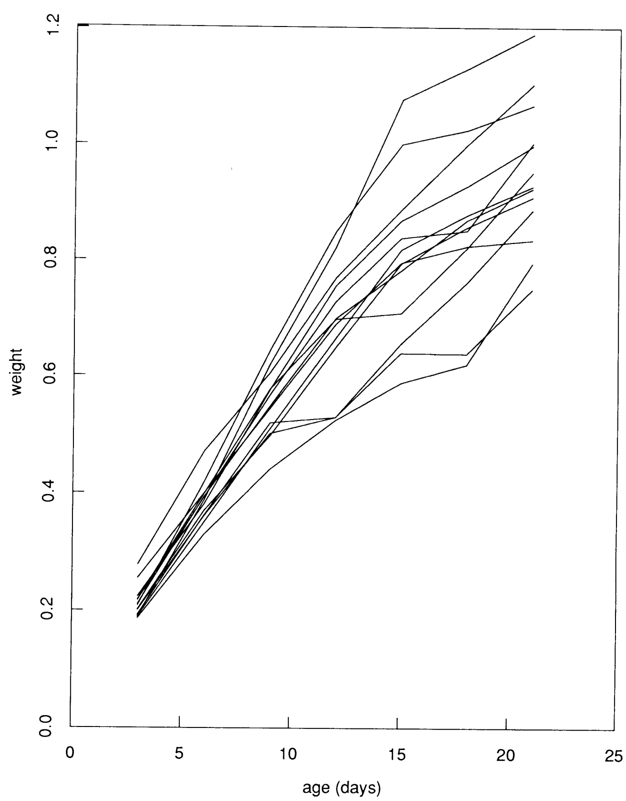

FIG. 1. *Superimposed individual growth curves for the mouse weight data set (the paper's Table 2).*

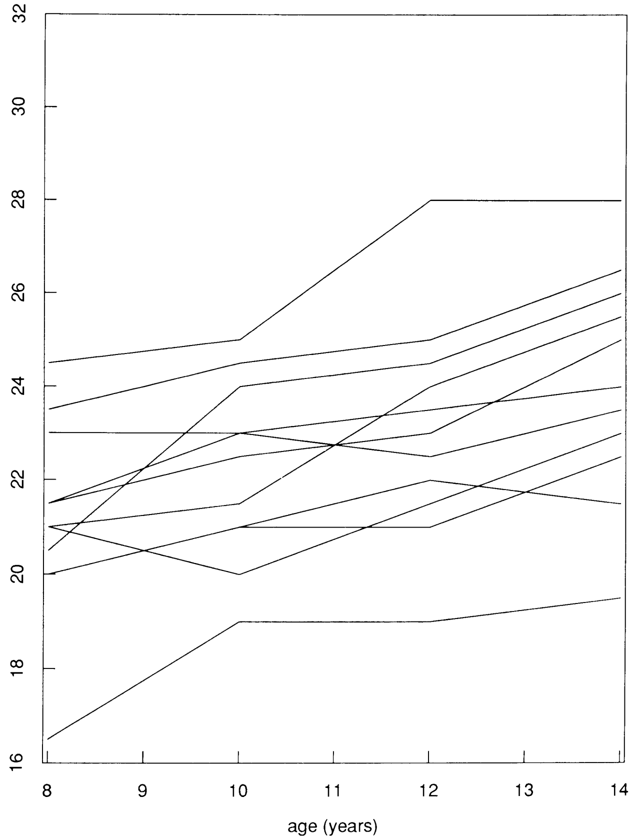

FIG. 2. *Superimposed individual growth curves for the* 11 *girls from the dental measurements data set* (*the paper's Table* 4).

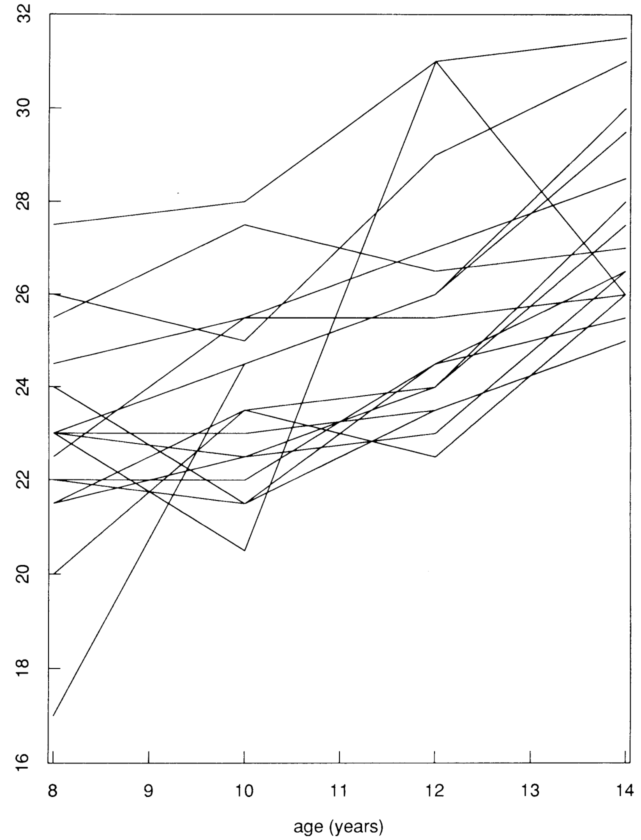

FIG. 3. *Superimposed individual growth curves for the* 16 *boys from the dental measurements data set* (*the paper's Table* 4).

("linear prediction") has nothing to do with growth curves, virtually ignoring the time ordering of the data, and some of the methods that *are* based on growth curves include models, like global linearity for all individuals, that such plots reveal clearly do not provide good predictions for the three data sets in the paper. As another example, the paper mentions in a footnote to its Table 4 that 11 of the dental measurement vectors in that table are from girls and 16 from boys, but nothing is done with this information, in spite of the fact that there are potentially important differences (Figures 2 and 3) between these two sets of growth curves: the boys exhibit systematically different departures from global nonlinearity than the girls do, and display nonmonotonicities over time at a rate about twice that of the girls. In addition, the two groups exhibit different variability around their basic growth curve shapes, suggesting that both the choice of growth curve families in the modeling and the cross-validation estimates of prediction error should be stratified on sex.

As a final remark along these lines, consider one of the growth curve plots, say Figure 1 on the mouse weight data, in light of the paper's "general finding $\cdots$ that much of the information for forecasting is contained in the immediate past few observations," particularly the conclusion (Section 4.1) concerning the "individual regression predictor" method that "the best procedure is to fit a straight line to just the two previous measurements $\cdots$ and extrapolate to predict $\cdots$" This conclusion is arrived at by fitting first global and then increasingly local low order polynomials to all individuals, and noting that the fit improves the more local its basis (see the paper's Table 6). To anyone starting at Figure 1, could this conclusion be anything more than an empirical rediscovery of Taylor's theorem?

There is another way of expressing the general point I am trying to make about the role of data in statistical work, which draws on the distinction between how economists and statisticians approach data. Conventional wisdom has it that economists start with theory and then look for data to see if the theory is any good, whereas statisticians start with data and look for models that seem to fit. This distinction has been the basis of a certain amount of name-calling (the distinguished President of The University of Chicago, Hanna Holborn Gray, once threw fuel on this fire in a speech on the triumph of models over data with the following description: "An economist is a person who sees something in the world around him, goes home, takes out paper and pencil, and tries to figure out if what he saw with his own eyes is possible"), but there is more at stake than style; this distinction has normative implications as well, as I tried to make clear above in my characterization of the four categories of data usage, and as I will attempt to illustrate below in the description of how I would approach Rao's modeling problem. It would seem that Rao has given us a fairly standard economics paper here, starting with some general theory about how prediction might be attempted and then (in his own words) "choos[ing three data sets] for illustration." I just wish he had gotten closer to his data.

2. ON EXCHANGEABILITY JUDGMENTS IN PREDICTIVE MODELING

The problem addressed by this paper is predictive in nature. Predictive modeling is the process of expressing one's beliefs about how the past and the future are connected. These connections are established through *exchangeability judgments*: with what aspects of past experience will the future be more or less interchangeable, after conditioning on relevant factors? It is not possible to avoid making such judgments; the only issue is whether they are made explicitly or implicitly. I claim that it is better to approach the modeling in a way that makes such judgments explicit.

Many of the methods in this paper do not do this. Most of its methods seem basically to be smoothing techniques, some simple, some elaborate, all sidestepping the issue of what beliefs the modeler has about how the past and the future are related. To illustrate this, the following is a brief restatement of the ideas behind a number of the paper's methods, for concreteness in the context of the mouse weight data (Table 2) and with the goal of predicting the weight of mouse 13 on day 21.

Linear Prediction

Regress y_{21} on $y_3, y_6, \cdots, y_{18}$ for individuals 1 to 12; then use this regression equation with the values of $y_3, \cdots, y_{18}$ for individual 13 to predict y_{21} for that individual. Do this over again with subsets of the predictors $y_3, \cdots, y_{18}$, using cross-validation to pick the predictor subset with the smallest mean squared prediction error. With this representation of the data, an entire growth curve is mapped onto a single point in $(p + 1)$-space, where p is the number of predictors, and the assumption is made implicitly that the individuals with which the current individual is more or less interchangeable are the ones that are close to him in the first p of these $p + 1$ coordinates. But this is a strange scale on which to be making exchangeability judgments, because no information about the time relationships among the predictor values is used. The regressor columns can be switched around to correspond to any time order desired, and this method will produce the same prediction. Potentially valuable derivative information—shape, rate of change of growth over time—is thus ignored, leading to the possibility of false judgments both of exchangeability and of nonexchangeability. In the latter category, for instance, consider two individuals with quite similar growth curves apart from a fairly large vertical shift (the first two boys in the ramus height data set, the paper's Table 3, are an example of this). On the "linear prediction" scale, these individuals will be far enough apart that there will not be any clear predictive linkage between them, whereas it would be obvious on the growth curve scale how to use information from the one to predict the other.

There is one way in which this method does pay attention to the time order of the data: when Rao does his variable selection by leaving out y_i's he drops the y's farthest back in time first. But, taken at face value, there is nothing in this "linear prediction" method that forces the oldest ones to go first—with this setup the regression doesn't know the meaning of "oldest." In fact, standard backward selection regression methods applied to the mouse data would suggest dropping the predictors not in the order $y_3, y_6, \cdots, y_{15}$ but in the order $y_{12}, y_9, y_{15}, y_3, y_6$.

Individual Regression Predictor

(1) Ignoring all other subjects, separately for each individual j (from 1 to 12) fit a polynomial in $(t_i; y_{ij})$ (in the notation of the paper's Table 1) of order k using the s measurements immediately preceding $t_7 =$ day 21, the time point to be predicted, and extrapolate this fit to predict. (2) Do cross-validation on mean squared prediction error over all 12 members j of the training sample to choose k and s. (3) When a new individual, mouse 13, comes along, use k and s from step (2) to fit the appropriate polynomial, again ignoring all other data, and extrapolate. With this method there is no attempt to form an exchangeability judgment on which member(s) of the training sample the new individual is most like; in fact, the only sense in which learning from past experience occurs is through the implicit position that whatever was good for the previous individuals as far as k and s is concerned must also be good for the new individual. This method performs predictably poorly on Rao's data sets (see his Table 8).

Calibration of Individual Predictors

As in the "individual regression" predictor approach, fit a polynomial of degree k to the s measurements preceding y_{21}, separately for each of the 12 individuals in the training sample, and extrapolate by using each of these polynomial fits to get a total of 12 predicted values $\hat{y}_{21}$. Form a new data set with 12 rows and two columns, the observed y_{21}'s and the predicted $\hat{y}_{21}$'s, and regress y_{21} on $\hat{y}_{21}$, obtaining $\tilde{y}_{21} = \hat{\beta}_0 + \hat{\beta}_1 \hat{y}_{21}$. When a new individual appears, fit a polynomial of degree k to the s previous measurements for him, extrapolate to get his $\hat{y}_{21}$, and plug into the regression equation. Cross-validate this whole process on mean squared prediction error to find the best k and s.

Regression on Polynomial Coefficients

Pick an individual j in the training sample consisting of the first 12 mice. Fit a kth degree polynomial to the s measurements preceding day 21, obtaining first stage regression coefficients $\hat{\beta}_{0j}, \cdots, \hat{\beta}_{kj}$. Do this for all $n = 12$ individuals and construct a new data set, using as columns y_{21} and the $\hat{\beta}_{ij}$. Regress y_{21} on the $\hat{\beta}$'s (?!), obtaining $k + 1$ second stage regression coefficients. When a new individual (mouse 13) comes along, fit a kth degree polynomial to the last s measurements for this individual, obtaining his first stage coefficients; then feed these into the prediction equation based on the second stage coefficients to get his $\hat{y}_{21}$. Cross-validate the overall process to pick k and s.

What can be said about the last two methods from the point of view of using past experience in the prediction of future events, other than that a kind of elaborate smoothing is taking place that obscures the role of any given individual's growth curve in the prediction of another individual's growth? The situation is even less clear in the paper's "Bayes," "empirical Bayes," mixed-model and factor analytic methods, although about the latter one comment is possible: factor analysis can be viewed in general as a kind of regression on fake predictor variables (the so-called "common factors"), which have to be estimated from the data along with the usual regression coefficients (the "factor loadings"). In Rao's problem, the predictor variables carry the information about the growth curve relationship between t and y_t. But why ignore the powerful information you have on these relationships, as conveyed by plots like Figures 1 to 3? No wonder the factor analytic methods don't do very well on Rao's data (see his Table 8).

Why is it better to approach predictive modeling in a way that makes the necessary exchangeability judgments explicit? After all, smoothing methods like those in the paper may work well enough to satisfy one's predictive accuracy goals in a given situation. There are three main reasons for preferring explicitness, one philosophical and two practical. On the practical side, because the smoothers do not make the necessary judgments explicitly, (1) if they *don't* predict accurately enough it is difficult to see how to improve them; and (2) you have no reason to believe that cross-validation uncertainty assessments based on such methods tell you anything about the predictive performance of those methods for future observations. But perhaps most importantly, on the philosophical side (Hodges and Draper, 1987), modeling can be viewed as a process of adding information to the data set, which in the absence of such added information refers only to itself. If we do not conduct the modeling activity in a way that makes our judgments explicit, how will we know when we are done what information we have added? How will we assess how much of the final answer is due to the insertion of modeling information unchallenged by the data and how much comes from the data speaking for itself?

What I Would Do Instead

My remarks have been too critical to follow Marcus Tullius Cicero's dictum about commenting on other people's work ("I criticize by creation, not by finding fault") to the letter, but in that spirit, at least, here is an outline of how I would approach Rao's prediction problem, in a way that attempts to make the exchangeability judgments explicit, to capture uncertainty in those judgments and to propagate that uncertainty through to the final predictions and uncertainty assessments. Time and space restrictions have prevented me from completing this analysis and presenting the results here; I hope to do so elsewhere.

Consider the individual growth curves of Figures 1 to 3 (which in the interests of saving journal space have been condensed into three plots, but which in practice should be examined separately). In a given data set, the methods in the paper under discussion all implicitly treat all individuals as interchangeble, when the growth curve plots indicate they may not be. Conceptually, one can imagine that each individual grows in a manner corresponding to one of a rather small number of *families* of growth curves, each indexed by a modest number of parameters. It is not necessary to know what it might be about the individuals structurally that divides them into subgroups, to improve the flexibility of one's modeling by positing a small number of such families of growth curves and then asking the data which subgroup each individual belongs to. In this formulation the central exchangeability judgment is the family to which a given individual's growth curve belongs; conditional on this choice, the data can help choose likely parameter values in the usual manner.

With this way of looking at the problem there will be three sources of uncertainty to assess: *structural* or *model* uncertainty about which growth curve family an individual belongs to, *estimation* uncertainty about parameter values conditional on the family and *prediction* uncertainty arising from the fact that, for each new individual, even if the family and the parameters were known, there would be stochastic fluctuation around the underlying growth curve. This characterization makes the modeling approach I have in mind a special case of a general process that might be called *propagation of model uncertainty* (Harrison and Stevens, 1971, 1976; de Finetti, 1974, 1975; Leamer, 1978; Smith, 1983; Hodges, 1987; Draper, Hodges, Leamer, Morris and Rubin, 1987). The idea is as follows.

In a predictive context, a model is just a joint probability distribution for the observables, so that one can conceive of the space of all possible models as a collection of such distributions. What we usually refer to as a "model," with unknown parameters, is a low dimensional curve in this space indexed by some parameters, and choosing a single "model" corresponds in a Bayesian sense to putting a prior distribution on model space that concentrates all its mass on this curve. (In what follows I will use "model" to denote a subspace of model space, indexed parametrically in this way.) When structural uncertainty is present, as it almost always is, such priors do not realistically reflect this uncertainty; to improve on usual practice it is necessary to entertain the possibility of a number of "models," by starting with a richer prior on model space that spreads it mass over more than one low dimensional curve.

This will make a new layer of integration necessary in the calculation of the posterior predictive distribution (ppd), $p(\text{future} \mid \text{data})$: in effect, one ends up mixing the ppd's conditional on each "model," $p(\text{future} \mid \text{data, "model"})$, to arrive at the overall ppd, by using as mixing weights the posterior probabilities of the "models" given the data, $p(\text{"model"} \mid \text{data})$. Symbolically,

$$p(\text{future} \mid \text{data}) = \int p(\text{future} \mid \text{data, "model"}) \cdot p(\text{"model"} \mid \text{data})\, d\,\text{"model."} \tag{2.1}$$

These "models" typically all have unknown parameters, so this process also involves an integration over uncertainty in the parameters, which has been suppressed in the above notation. The necessary ingredients turn out to be a prior on model space, $p(\text{"model"})$, priors on the parameters conditional on the "model," $p(\text{parameters} \mid \text{"model"})$, likelihoods for each "model," $p(\text{data} \mid \text{"model," parameters})$ and predictive distributions for each "model" given the data and parameters, $p(\text{future} \mid \text{data, "model," parameters})$. This may seem like a lot to require, particularly when other approaches to predictive modeling, like those of the paper under discussion, appear to avoid having to specify such distributions. But choices of this type must be made; the only issue is whether or not to make them explicitly. See Leamer (1978), Smith (1983) and Draper, Hodges, Leamer, Morris and Rubin (1987) for further details.

More specifically, in the context of the mouse weight data, here is how the modeling would go. If covariates were present, as in the dental data, the first step would be to think about how such information should be used to condition your exchangeability judgments, for instance by carrying out the steps below separately for girls and boys. The decision, if it turns out that way, that the exchangeability need not be conditional on sex would then be made only after the data support this choice.

(1) *Exploratory analysis to choose* {*prior on model space*} = {*choice of growth curve families*}. Draw separate curves for all individuals in the training sample and study them. Examination may well suggest two or three basic growth curve families (global linearity, general downward curvature and basic sigmoidal shape, for example), each indexed by a small number of parameters (in the mouse data, not more than three); for concreteness, say three families, with 2, 3 and 3 parameters, respectively.

(2) *Mixing models.* Starting with a flat prior (1/3, 1/3, 1/3) on the three curve families, for lack of an initial reason to differentiate between them as more and less probable, and essentially flat priors on the parameters within these families, pick an individual in the training sample and update to a posterior distribution on model space, $p(\text{curve family} \mid \text{data})$. Do this separately for each individual in the training sample and look at all the resulting three vectors. If you have done a good job choosing the families of growth curves, most of these vectors will have dominant components, for example (0.8, 0.05, 0.15) or (0.22, 0.74, 0.04). If there are a number that are close to (1/3, 1/3, 1/3) (the standard deviations of each of these vectors could be used as a numerical index of closeness to a flat posterior, somewhat like the varimax criterion in factor analysis), you probably don't have a rich enough set of growth curve families yet (for those individuals, the data are trying to say that none of the curves fits terrifically). If this happens, go back to the prior on model space step and consider enlarging the number of growth curve families you're willing to propagate uncertainty over, repeating these first two steps as necessary. Eventually, when you end up with a satisfactory set of growth curve families, start over again with a flat prior on model space and on the parameters conditional on the "models," take all the training sample individuals as a batch and update to posteriors on model space and on the parameters conditional on the curve families.

(3) *Prediction for the new individual.* When a new individual appears, using the posteriors on model space and the parameters conditional on the "models"

from the training sample as priors, construct a posterior predictive distribution for the unobserved time point given that individual's growth curve up to that point. The result of all this will be an analysis in which the data help make the necessary exchangeability judgments adaptively, and in which the posterior predictive variability captures all three sources of uncertainty above—structural, estimation and prediction.

I am grateful to Professor Rao for having written a paper that provoked a great deal of thought in me, and I look forward to comparing the results of this propagation of uncertainty analysis with those from his prediction methods and from other approaches to prediction in growth curve models.

ACKNOWLEDGMENTS

In preparing these remarks I have had the benefit of comments from James Hodges and John Rolph.

ADDITIONAL REFERENCES

ANDREWS, D. F. and HERZBERG, A. M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker.* Springer, New York.

BRILLINGER, D. R. (1986). Comment on "Generalized additive models," by T. Hastie and R. Tibshirani. *Statist. Sci.* **1** 310–312.

DE FINETTI, B. (1974, 1975). *Theory of Probability,* **1**, **2**. Wiley, New York.

DEGROOT, M. H. (1987). A conversation with C. R. Rao. *Statist. Sci.* **2** 53–67.

DRAPER, D., HODGES, J. S., LEAMER, E. E., MORRIS, C. N. and RUBIN, D. B. (1987). A research agenda for assessment and propagation of model uncertainty. The Rand Corporation, N-2683-RC, November 1987.

DUMOUCHEL, W. H. and HARRIS, J. E. (1983). Bayes methods for combining the results of cancer studies in humans and other species (with discussion). *J. Amer. Statist. Assoc.* **78** 293–315.

HARRISON, P. J. and STEVENS, C. F. (1971). A Bayesian approach to short-term forecasting. *Oper. Res. Quart.* **22** 341–362.

HARRISON, P. J. and STEVENS, C. F. (1976). Bayesian forecasting (with discussion). *J. Roy. Statist. Soc. Ser. B* **38** 205–247.

HOAGLIN, D. C., MOSTELLER, F. and TUKEY, J. W., eds. (1983). *Understanding Robust and Exploratory Data Analysis.* Wiley, New York.

HOAGLIN, D. C., MOSTELLER, F. and TUKEY, J. W., eds. (1985). *Exploring Data Tables, Trends, and Shapes.* Wiley, New York.

HODGES, J. S. (1987). Uncertainty, policy analysis and statistics (with discussion). *Statist. Sci.* **2** 259–291.

HODGES, J. S. and DRAPER, D. (1987). On the information content of models. In preparation.

LAGAKOS, S. W., WESSEN, B. J. and ZELEN, M. (1986). An analysis of contaminated well water and health effects in Woburn, Massachusetts (with discussion). *J. Amer. Statist. Assoc.* **81** 583–614.

LEAMER, E. E. (1978). *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* Wiley, New York.

MOSTELLER, F. and WALLACE, D. L. (1984). *Applied Bayesian and Classical Inference: The Case of the Federalist Papers.* Springer, New York.

REINSEL, G. C. and TIAO, G. C. (1987). Impact of chlorofluoromethanes on stratospheric ozone. *J. Amer. Statist. Assoc.* **82** 20–30.

SEARLE, S. R. (1971). *Linear Models.* Wiley, New York.

SMITH, A. F. M. (1983). Bayesian approaches to outliers and robustness. In *Specifying Statistical Models: From Parametric to Nonparametric, Using Bayesian and Non-Bayesian Approaches* (J. P. Florens, M. Mouchart, J. P. Raoult, L. Simar, and A. F. M. Smith, eds.). Springer, New York.

VELLEMAN, P. F. and HOAGLIN, D. C. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis.* Duxbury Press, Boston.

WEISBERG, S. (1986). A linear model approach to backcalculation of fish length. *J. Amer. Statist. Assoc.* **81** 922–929.

Comment

Alan Julian Izenman

1. INTRODUCTION

It gives me great pleasure to comment on this paper by Professor Rao. The central issues raised here are choice of a prediction model and assessment of associated prediction errors for growth curve data. Professor Rao has given us a number of different approaches to these problems. I offer a few general comments and some specific comments, mention alternative directions in growth curve modeling and prediction and also make some comments on the mice data used in this paper.

Alan Julian Izenman is Associate Professor, Department of Statistics, Temple University, Philadelphia, Pennsylvania 19122.

2. SOME GENERAL COMMENTS

For a statistician, context should always play a role in the modeling process. Too often, data are analyzed without regard to the original purpose of their collection. This can be especially true when modeling a growth process where biological reasoning may help in the modeling and subsequent interpretation of results. The first thing I noticed about this paper is that there is no clear description of the three data sets used as illustrations or why they are even interesting for prediction purposes. (Does anyone understand what