lar treatment and the points at which the curve bifurcates are precisely those at which the treatments begin to differ. Again, it would be hard to estimate such a bifurcating curve using kernel methods, but a way of thinking based on penalized least squares and spline smoothing gives a natural way to proceed.

In conclusion, it would be facile to suggest that any particular method will yield answers to all the problems one might encounter. But I do believe that if we are making detailed comparisons between different approaches, the spline smoothing method as a general approach has a great deal to commend it.

Rejoinder

C.-K. Chu and J. S. Marron

We are very grateful to the discussants for their interesting and thoughtful additions to the points we made in this paper. We also thank the editor for many helpful comments, and for the nontrivial task of organizing the discussion.

Our response is organized into sections, with the first three concerning topics raised by more than one discussant, followed by some individual responses in alphabetical order.

1. PHILOSOPHICAL ISSUES

Useful elaboration of our P1-P2 formulation of the viewpoints that have been adopted to consider smoothing, has been provided by Grund and Härdle and by Silverman.

We agree with Grund and Härdle that computational issues are very important and welcome the addition of their P3 as a general principle. However, in the present context, we do not view proper incorporation of this factor as having a major impact on the ideas indicated here. The reason is that both $\hat{m}_C$ and $\hat{m}_E$, when properly implemented, for example, as described in Section 3 of Grund and Härdle, have roughly comparable computation time. On the other hand, this P3 could easily become vital in, for example, a comparison of splines versus kernels as suggested by Silverman.

We also find Silverman's P4 and the surrounding discussion very useful. This is a very nice extension of the points we were trying to make. One small point we would like to clear up is that when we attached the phrase "nonparametric regression estimation" to P2, we were referring to the phrase, not the methodology. Our intention was to convey the point that most people who have used this phrase in the literature tend to lie in the P2 camp. However, we wholeheartedly endorse Silverman's main point (also expressed well by Grund and Härdle) that there needs to be more combined use of P1 and P2.

2. ADDITIONAL COMPARISON

Some interesting and carefully considered alternative ways of comparing $\hat{m}_C$ and $\hat{m}_E$ are presented by Grund and Härdle and by Hart. We welcome these deeper analyses and are happy that the main conclusions are not much different from what we saw by simpler methods.

The figures of Grund and Härdle are very informative and provide excellent visual quantification of the points we were driving at in the paper. However, we caution against trying to infer too much from these examples. We are hesitant to make a recommendation as to which estimator is better on the basis of the size of the region where the ratio is bigger than one, because this is only one example. Even if one looked at several such examples, there are doubtless other examples that give the opposite conclusion. Furthermore, even in the presented example caution is indicated, because these sizes of regions are also dependent on the parameterization that has been chosen for the example. For example, the regions seen in Figure 4 could look quite different if this picture were based on the logarithms of these two parameters. But, of course, the main point here is that $\hat{m}_C$ and $\hat{m}_E$ are not really comparable in terms of one being always preferable, and the figure illustrates this in a compelling fashion.

Hart's idea of looking at the joint probability structure of $\hat{m}_C$ and $\hat{m}_E$ is excellent. He admirably illustrates that this is an important issue, and things are not as one might expect at first guess. This clearly needs to be borne in mind in future comparisons of estimators.

3. OTHER KERNEL ESTIMATORS

Hart and Jones discuss alternative kernel smoothers to $\hat{m}_C$ and $\hat{m}_E$ and make some interesting cases for their serious consideration. We were

aware of these others, but chose to focus only on $\hat{m}_C$ and $\hat{m}_E$ because they are the two most often featured in the current literature. Our intuitive feeling about the Priestley–Chao estimator had been that it was "roughly the same as $\hat{m}_C$, but sometimes a bit worse," although Jones makes it clear that this idea needs some careful reconsideration. We ruled out the Yang estimator brought up by Hart on the grounds that once we opened the door to what we viewed at the time as "obscure alternative methods," we then needed to also consider a host of other possibilities, such as nearest neighbor estimators, local polynomials and a large family of hybrids. Our feeling was that consideration of too many estimators would have obscured the main ideas. This may have been a mistake in view of the nice points made by these discussants, but it seems best at this point to put this topic into the category of future work to be done.

4. INDIVIDUAL RESPONSES

Gasser, Jennen-Steinmetz and Engel

This discussion is critical of our use of the word "efficiency" in Section 3. Of course, we agree that MSE should be the central issue here, and not variance alone. However, note that MSE is the sum of variance and squared bias, so it can be well understood by analyzing each of these in turn, which we did in Sections 3 and 4, respectively. In Section 3, we carefully chose some settings where the bias was essentially the same for the two estimators. Hence, differences in MSE are directly determined by differences in variance, so our use of the phrase efficiency seems justified. However, we do agree that "Variance Issues" may have been a better title for Section 3. We agree that a more careful analysis than ours would look at error, where each estimator uses its best possible bandwidth. However, the analyses of Grund and Härdle and of Hart show that our analysis did succeed in conveying the important ideas.

Gasser and others were also critical of our occasional realization-wise study. While we agree that expectation should play an important role, we submit that it is very useful to consider individual realizations as well. Both viewpoints are important, and we feel it is not enough to restrict attention to either one alone. We believe the insights that we pointed out in the paper make this clear, but note also Silverman's excellent additional points about the attractive feature of "continuity of estimators" that is satisfied by $\hat{m}_E$ but not $\hat{m}_C$, which is certainly a realization-wise property.

These discussants also objected to some of our examples, especially the contrived ones. Our general feeling on this topic is that main points are best supported through a variety of methods. While one should never rely completely on artificial examples, they can aid in the presentation of ideas. We consider them to be quite useful as illustrative tools, especially when used in combination with other methods, such as real data examples and asymptotic analysis as we have done here. For example, the variance comparison in our (3.1) and (3.2) has been known for some time, but we believe that the first real intuitive understanding of what drives this difference is provided by our Figures 4, 5, 6 and 7 and the surrounding discussion. Note that this careful realization-wise study has given a clear understanding of the forces behind what can be seen in the expectation sense.

We generally agree with the points made in the Bias part of this discussion, and consider them a worthwhile amplification of what we had to say in our Section 4. Their Figure 1, which also appeared in Gasser and Engel (1990), is an important and clever addition to the examples in our Figure 11. The one controversial point here is what is the "natural" way to handle random design data. We respect the discussant's view on this, but stress that much different ones have been considered reasonable as well. Our personal feelings on this are more along the lines indicated in the discussion of Jones.

A very good point is made here about boundary effects. It is true that our presentation unfairly penalizes $\hat{m}_C$ by not doing boundary adjustments (since it has a much stronger need for these). We did consider doing boundary adjustments in our examples, but, mostly from laziness, gave up when it became clear it would take a large amount of effort to implement the complicated methods (note that this is our personal view, for instance Hart expresses quite a different one) that are usually suggested. This seems partially justified by the fact that such adjustments are also unappealing because they detract from what we view as the major strengths of the kernel method: simplicity and interpretability. We do not agree with the discussants that there are no methods for boundary adjusting $\hat{m}_E$ (in the equally spaced fixed design case). See Rice (1984) and also two unpublished proposals involving "data reflection" by Wu and Chu (1990) and by Hall and Wehrly (1991).

These discussants make some good points about convenience of construction of plug-in bandwidth selectors. We do not see for sure that the bias of $\hat{m}_E$ is harder to estimate, but algorithms of this type will certainly be quite a lot more complicated. However, there are important competitors to the plug-in approach, such as the double smoothing idea discussed recently in Härdle, Hall and Marron (1991),

for which the two methods are quite comparable in this sense.

Gasser and others take issue with our skepticism concerning the Gasser and Engel (1990) minimax result, but only choose to address the first of our three reasons. Their criticism of this point is based on the fact that we here allow f to take on values close to 0, but not in our assumption (A.4). We would like to repeat that the case of f close to 0 in some locations should not be ruled out, as it includes important situations such as Gaussian f. We see no irony in our assumption (A.4), as *no attempt* is made at that point to be general. As indicated before assumption (A.1), all of these assumptions have been chosen for simplicity and clarity of presentation. They can be weakened a lot, but the added complication in terms of presentation seems unjustified in the present paper. We find it clear that our *main ideas* will still hold up for f Gaussian, but this is not the case for the Gasser and Engel minimax result.

Finally, we would like to clear the air concerning quotes. We apologize for the unnecessarily inflammatory construction of our statement, to which they have objected. The main point here is that Gasser and Engel place greater importance on bias than we do in the necessarily personal trade off between bias and variance. This became clear to us from their statement, "For random design convolution estimators have to pay a price in terms of variance. This can be better tolerated since replications of studies can control for random phenomena." While our own view on how this trade off should be made is different, we respect theirs. We also hope this part of the discussion does not detract from their main point, which is the quote stated in the discussion.

Jones

This discussion brings up an interesting general framework for classifying kernel estimators, in terms of *internal* and *external*, which should be quite useful for studies going beyond this one into the many other possible types of smoothers.

There is a good point here about calling $\hat{m}_C$ the "convolution" estimator, because $\hat{m}_E$ can also be viewed as a type of discrete convolution. However, the first thing that jumps to our minds is the simple continuous convolution, which $\hat{m}_C$ most clearly is. Another justification might be the very simple convolution form of the bias, for $\hat{m}_C$, relative to $\hat{m}_E$.

Silverman

Among the many interesting points in this discussion is a nice motivation for consideration of smoothers that do not fall exactly into the kernel framework. Silverman (and a number of other important researchers as well) has a personal preference for the smoothing spline, but there are other smoothing methods that also have strong advocates, and also deserve mention here, including B-splines (also sometimes called regression splines, and much different from smoothing splines), orthogonal series and wavelet estimators. Proper comparison of these methodologies is clearly beyond the scope of the discussion here, but we can not resist adding a few comments.

Silverman's point about splines providing simple and appealing approaches to special problems is well taken. Actually, the second author of this paper first heard this view expressed some years ago by Doug Nychka. Nychka made the point especially dramatically at a conference, where the material in the paper by Silverman and Wood (1987) was presented in a talk by Silverman. At the end of the talk, there was time for questions, and Nychka said, "I have a question, not for the speaker, but for Steve Marron: Can he think of a kernel approach to this problem?" Marron could not think of one on the spot, and further attempts later only gave solutions that were unsatisfactory because they were too complicated. Definitely the clever ideas of Silverman and Wood worked very well in this case, and it should be stressed that Silverman's example is only one of many we have seen.

However, despite the attractive features of smoothing splines (and the other estimators as well), and the fact that there have been many superb data analyses carried out with these, we are skeptical that they can ever replace the leadership role of kernel estimators. In our opinion, the reason that kernel estimators have been and will continue to be so popular, despite their acknowledged drawbacks, is that they are so simple. An important benefit of this is ease of implementation, but we believe the biggest payoff for this simplicity comes in terms of interpretability. Even people with little education at all, to say nothing of the nontrivial mathematical background required to understand a spline, can grasp the visual idea of a moving average, and understand intuitively the structure that they see in such a curve. However, even strong spline advocates themselves have trouble explaining just what their estimator is doing to the data. When queried on this issue, their first response is to point out that it is the solution to the appealing minimization problem underlying smoothing splines. When pressed by inquirers who are not satisfied that they understand what is being done to the data, the next response is: "Silverman (1984) showed that the spline is often behaving much like a kernel estimator." Of course, the natural next

question is, "If your insight comes from kernel estimation, shouldn't that be what you are using to analyze your data?"

While we have presented here an alternative viewpoint to Silverman's, we should point that this is far from the last word on the subject. For example, Silverman points out some nice properties of the spline, not naturally shared by kernel estimators, that come from his approximation by a rather special type of kernel estimator.

In summary, we doubt that either type of estimator will put the other out of business. As with the comparison of $\hat{m}_C$ and $\hat{m}_E$, both kernels and splines have their relative merits, and it seems clear to us that both will continue to attract adherents.

5. CONCLUSIONS

We would like to again thank the discussants for their fine contributions. While many different opinions have been expressed here, we believe we can speak for all in saying that nonparametric regression estimation methods (i.e., scatter plot smoothers) have been abundantly demonstrated to provide powerful data analytic tools. Furthermore, we think that their future is very bright, because scientists are recently becoming very ambitious in terms of attempting to gain insights from more and more complex data sets, often with relatively less and less a priori information available as to which models are appropriate.

ADDITIONAL REFERENCES

BREUER, K. (1990). *Approximation von Kernglättern durch die WARPing-Methode.* Diplomarbeit, Fachbereich Statistik, Universität Dortmund.

CARROLL, R. J. and HÄRDLE, W. (1989). Symmetrized nearest neighbor regression estimates. *Statist. Probab. Lett.* **7** 315-318.

GASSER, T., KNEIP, A. and KÖHLER W. (1991). A flexible and fast method for automatic smoothing. *J. Amer. Statist. Assoc.* **86** 643-652.

GREEN, P. J. (1987). Penalized likelihood for general semiparametric regression models. *Internat. Statist. Rev.* **55** 245-259.

HALL, P. and WEHRLY, T. E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *J. Amer. Statist. Assoc.* **86** 665-672.

HÄRDLE, W. (1991). *Smoothing Techniques with Implementation in S.* Springer, New York.

HÄRDLE, W., HALL, P. and MARRON, J. S. (1991). Regression smoothing parameter selectors that are not far from their optimum. *J. Amer. Statist. Assoc.* To appear.

HART, J. D. and WEHRLY, T. E. (1991). Kernel regression when the boundary region is large, with an application to testing the adequacy of polynomial models. *J. Amer. Statist. Assoc.* To appear.

HASTIE, T. and TIBSHIRANI, R. (1990). *Generalized Additive Models.* Chapman and Hall, London.

JOHNSTON, G. J. (1979). Smooth nonparametric regression analysis. Univ. North Carolina, Inst. Statistics, Mimeo Series 1253.

JONES, M. C. (1991). On correcting for variance inflation in kernel density estimation. *Comput. Statist. Data Anal.* **11** 3-15.

JONES, M. C. and DAVIES, S. J. (1991). Versions of the kernel regression estimator. Preprint.

MACK, Y. P. and MÜLLER, H.-G. (1989b). Derivative estimation in non-parametric regression with random predictor variable. *Sankyā Ser. A* **51** 59-72.

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models.* Chapman and Hall, London.

MÜLLER, H. G. (1987). Weighted local regression and kernel methods for nonparametric curve fitting. *J. Amer. Statist. Assoc.* **82** 231-238.

PARZEN, E. (1981). Nonparametric statistical data science: A unified approach based on density estimation and testing for "white noise." Technical Report, Dept. Statistics, Texas A&M Univ.

PRIESTLEY, M. B. and CHAO, M. T. (1972). Non-parametric function fitting. *J. Roy. Statist. Soc. Ser. B* **34** 385-392.

SILVERMAN, B. W. (1978). Density ratios, empirical likelihood and cot death. *J. Roy. Statist. Soc. Ser. C* **27** 26-33.

SILVERMAN, B. W. (1984). Spline smoothing: The equivalent variable kernel method. *Ann. Statist.* **12** 898-916.

SILVERMAN, B. W. (1985). Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). *J. Roy. Statist. Soc. Ser. B* **47** 1-52.

SILVERMAN, B. W. and WOOD, J. T. (1987). The nonparametric estimation of branching curves. *J. Amer. Statist. Assoc.* **82** 551-558.

STUTE, W. (1984). Asymptotic normality of nearest neighbor regression function estimates. *Ann. Statist.* **12** 917-926.

TUKEY, J. W. (1977). *Exploratory Data Analysis.* Addison-Wesley, Reading, Mass.

WU, J. S. and CHU, C.-K. (1990). Kernel estimators with projected data in nonparametric regression. Unpublished paper, Inst. Statistics, National Tsing Hua Univ.

WU, J. S. and CHU, C.-K. (1991). Double smoothing for kernel estimators in nonparametric regression. Preprint.

YANG, S. S. (1981). Linear functions of concomitants of order statistics with applications to nonparametric estimation of a regression function. *J. Amer. Statist. Assoc.* **76** 658-662.