Rejoinder

M. P. WAND, J. S. MARRON, and D. RUPPERT

We are grateful to the discussants for their interesting comments and suggestions. It is our belief that several very important issues in practical density estimation have been addressed in this discussion.

The idea of transforming to the uniform distribution is intriguing. Rudemo has suggested transforming by a parametrically estimated cdf. This strategy replaces the question "What family of transformations is appropriate for my data?" by the question "What parametric model is appropriate?" Statisticians may feel more comfortable with addressing the second question. Rudemo further suggests estimating the parameters by maximum likelihood. This would be easier than our method of minimizing an estimate of MISE_Y, but raises questions of robustness since the parametric family is considered to be only a rough approximation to the actual density of the data. Another potential drawback is that maximum likelihood is strongly connected to Kullback–Leibler distance, which has been shown by Hall (1987) to have unattractive properties for curve estimation, because it is too sensitive to tail behavior. Since the ultimate goal is to get a good nonparametric density estimate, we feel more comfortable minimizing an estimate of MISE rather than maximizing a likelihood.

Recent work of Ruppert and Cline (1990) takes a slightly different approach—transformation by a nonparametrically estimated cdf. Their transformation is a smooth kernel estimate of the cdf. The smoothness is essential. Ruppert and Cline showed that the effect of transformation is to reduce the order of the bias. The result is that one can use a larger bandwidth and, therefore, that rates of convergence equal to those of higher-order kernels are achievable, but with an estimate guaranteed to be nonnegative. At the time the current article was written, we saw transformation as a means of varying the amount of smoothing across the range of X. However, transformation does more. It also changes the amount of bias for a given effective bandwidth. This is most pronounced for transformations to the uniform distribution but occurs for any nonlinear transformation. Since our method for choosing the bandwidth and the transformation parameters minimizes an estimate of MISE, it has the effect of both minimizing bias over the family of transformations and optimizing the bias–variance trade-off inherent in the choice of bandwidth.

The problems with estimating a shift parameter by maximum likelihood are very nicely discussed by Atkinson (1985). The grouped likelihood approach of Atkinson, Pericchi, and Smith (in press) seems very clever and promising.

Certain alternatives to maximum likelihood do not have the difficulties that Atkinson discusses. For example, in the context of analysis of variance, Berry (1987) estimated the shift of the shifted log transformation by minimizing either the absolute value of the skewness of the residuals, or their absolute kurtosis (where kurtosis is the standardized fourth moment minus 3), or the sum of these two measures of deviation from normality. Working with the transform-both-sides model of Carroll and Ruppert (1984, 1988), Nakamura and Ruppert (1990) showed how to estimate both parameters of the shifted power transformation family. They studied both the extension of Berry's estimator to two parameters and an estimator based on a general test for symmetry applied to the residuals. The work of Nakamura and Ruppert uses the reparameterization developed in the present article and should clarify what we were recommending in our final paragraph.

Duan raises concerns about the sensitivity of transform–retransform methods to minor perturbations of the transformations. Working with essentially the same prediction model as Duan (1983) and Taylor (1986), Carroll and Ruppert (1988, in press) studied interval prediction of a new response, not just its mean. They found that the prediction intervals were quite sensitive to minor changes in the transformation parameters. This seems to be an accurate reflection of uncertainty about the variance and skewness of the new response, rather than a problem with the validity of the method.

We experimented extensively with the transformation parameters used on the income and suicide data. Our impression is that the density estimates are rather insensitive to minor parameter transformations when the d_1 and d_3 parameterization is used.

The ripples at the left in Figures 1b and 4b can be eliminated by increasing the shift parameter. This would change the extreme left of Figure 1b to look like that of Figure 1a with $h = .033$, but otherwise Figure 1b would be essentially unchanged. We agree with Scott that our minimum-MISE method of bandwidth selection does not appear to "feel" these ripples since they represent such a small fraction of the data. MISE may feel how rapidly the estimate rises from near 0 to the first peak, since the rise is slightly steeper in Figure 1b than in Figure 1a, with $h = .033$. Estimated MISE is not a sacred cow, and in this situation we would be comfortable modifying the transformation to eliminate the ripples.

We disagree with Duan's suspicion that the "adaptive kernel estimate is the only viable tool" (p. 355) for dealing with moderately skewed data. The purpose behind introducing the shift parameter into the power transformation family, as well as changing to the d_1 and d_3 parameterization, was to handle all types of skewness, in particular, differing amounts on the left and right sides.

It is our feeling that the transformation kernel estimator

Journal of the American Statistical Association
June 1991, Vol. 86, No. 414, Theory and Methods

has the potential to effectively estimate any feature that can be handled with an adaptive kernel estimator—especially with the current development of a wide variety of parametric as well as nonparametric approaches to the transformation. When we wrote the current article, we thought that the adaptive kernel estimator could not be computed quickly using binning methodology. We are very grateful to Scott for demonstrating that this is not the case, and we believe that his algorithm makes a very important contribution to the field. Scott's comments on the spurious behavior of an adaptive kernel estimator for certain sample sizes appear to be cause for concern, however.

Müller and Zhou demonstrate that the local bandwidth density estimates are more flexible than globally smoothed estimates. Although this is certainly true, we note that a trade-off needs to be made between flexibility and complexity (similar to the parametric versus nonparametric issue). These discussants express concerns that our automatic selection algorithm is complex, with which we do have some concurrence. However, we also question the complexity and, especially, the computational burden of computing automatically generated local bandwidth estimators for large samples.

Transformation/kernel estimators and variable bandwidth estimators are both undergoing vigorous development. We expect them to remain close competitors. We wonder if there might be some advantage to combining them.

Müller and Zhou's examples nicely make the point that estimation near the boundaries is difficult. We agree that there are clear limitations as to how much information is available in the data. Although boundary kernels with negative weights are useful, we have observed density estimates taking on negative values, when using the simpler boundary kernels of Rice (1984). This can be avoided with a "mirror image" or "boundary folded" type of adjustment, but there is a heavy price in terms of increased bias. We are currently working on a method, which eliminates this extra bias and is always nonnegative, by a transformation such that $f'_Y = 0$ at the boundary.

Rudemo's idea of applying the transformation technique to the histogram is very appealing, particularly the point about the bin widths relecting the effect of the transformation. It should be noted that direct application of this idea will, in general, yield an estimate that is not a step function. It seems straightforward, however, to replace this estimate by an approximant that is a step function. Furthermore, automatic "plug-in" implementation of this estimator would require estimation of $\int (f'_Y)^2$. Scott and Terrell (1987) provided some useful methodology for accomplishing this.

Last, we agree with Duan that logograms are a useful tool for checking normality. For data analysis in general, however, we prefer density estimates, because they most effectively show the distribution of probability mass. A logogram of the income data, for example, would not seem as informative as the density estimate in Figure 1b.

ADDITIONAL REFERENCES

Berry, D. A. (1987), "Logarithmic Transformations in ANOVA," *Biometrics*, 43, 439–456.

Carroll, R. J., and Ruppert, D. (1984), "Power Transformations When Fitting Theoretical Models to Data," *Journal of the American Statistical Association*, 79, 321–328.

——— (in press), "Prediction and Tolerance Intervals With Transformation and/or Weighting," *Technometrics*.

Hall, P. (1987), "On Kullback–Leibler Loss and Density Estimation," *The Annals of Statistics*, 15, 1491–1519.

Nakamura, M., and Ruppert, D. (1990), "Semiparametric Estimation of Symmetrizing Transformations With Application to the Shifted Power Transformation," unpublished manuscript.

Rice, J. (1984), "Boundary Modification for Kernel Regression," *Communications in Statistics, Part A—Theory and Methods*, 13, 893–900.

Ruppert, D., and Cline D. B. H. (1990), "Transformation-Kernel Density Estimation: Bias Reduction by Empirical Transformations," unpublished manuscript.