A personal view of smoothing and statistics

J. S. Marron

Department of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260, USA

Summary

Personal views are offered on the foundations of smoothing methods in statistics. Key points are that smoothing is a useful tool in data analysis, and that the field of data based smoothing parameter selection has matured to the point that effective methods are ready for use as defaults in software packages. A broader lesson available from the discussion is: a combination of computational methods and mathematical statistics is a powerful research tool.

Keywords: bandwidth selection, kernel estimation, nonparametric curve estimation, smoothing

1 Introduction

This paper contains some personal views on smoothing and its role in statistics. These views have been shaped by interaction with many people having very diverse opinions. These people are both active methodological statisticians, and also interested non-statisticians. The statisticians include "mainstream smoothers", but many others as well. An important component of my views comes from investigation of the fact that there are many very divergent, yet strongly held, opinions of "smoothing insiders" on issues such as what is the "right" choice of smoothing method.

Section 2 contains what I have learned about choice of smoothing method. Different people have made, and will continue to make, different choices. The main point is that the great majority of choices made are sensible, because there are many noncomparable factors involved in this choice. Diverse personal weightings of these factors result in choices that are different, but equally valid when viewed from an overall perspective.

Section 3 discusses modern bandwidth selection. The main lesson is that recent research has resulted in methods that are quite effective in finding "the right amount of smoothing". They are far better than the current defaults in most software packages, and in the case of kernel density estimation are ready for use in that role.

The discussion of research in bandwidth selection illustrates some important ideas on methodological statistics in general. In Section 4 it is suggested that a combination of mathematical and computational tools provides a particularly powerful approach.

A final point is: "smoothing is a useful method of analyzing data". Reasons for not losing sight of this point are given in Section 5.

2 Choice of smoothing method

A frequent occurrence is that a scientist consults a statistician (or a non-smoothing statistician goes to a "smoother") and says "I have heard about this smoothing stuff and would like to try it, how do I do it?" A number of different answers could be given depending largely upon whom is asked. The recommended method could be any of:

M1 a kernel - local polynomial method.

M2 a smoothing spline.

M3 a regression - B spline (much different from b).

M4 a Fourier smoother.

M5 a wavelet method.

M6 LO(W)ESS.

All of these methods have their firm adherents, who enjoy espousing the strengths of their personal favorite. It is interesting that the answers can be so different. Much personal discussion with involved parties has yielded the personal conclusion that everybody is right in their different choices. This seemingly contradictory statement is possible because there are several *noncomparable* factors involved. In particular, factors that go into the choice of smoother include:

F1 availability (is it "right there" in your software package?).

F2 interpretability (what does the smooth tell us about the data?).

F3 statistical efficiency (how close is the smooth to "the true curve"?)

F4 quick computability.

F5 integrability into general frameworks (e.g. S/S+).

F6 ease of mathematical analysis.

All of the methods M1 - M6 listed above have differing strengths and weaknesses in these divergent senses. None of these methods dominates any other in all of the senses F1 - F6. Since these factors are so different, almost any method can be "best", simply by an appropriate personal weighting of the various factors involved. Claims of "wretched performance" are also valid, for any of the above methods, when one adopts a suitably narrow viewpoint.

This is not the place to give a complete listing of the relative strengths and weaknesses of M1 - M6, but I shall discuss one case in more detail. The simplest possible naive fixed bandwidth local constant kernel methods have met with substantial criticism. This is justified, because they are relatively weak with respect to F3. I suspect that a contributing factor to this criticism is the popularity of such methods in the theoretical literature, much of which is due to strength in F6. But from a broader viewpoint, this criticism does not provide sufficient reason to rule out their use. In particular, these methods are the best (in my opinion) at F2. Fixed bandwidth local constant kernel methods put the interested analyst in closest possible intuitive contact with the data, because they are simple, understandable, local averages. Note that I am not advocating this estimator as the solution to all problems (in fact it has its drawbacks as does every method, and there are situations where other methods will be much more desirable), but instead am merely pointing out it cannot be dismissed out of hand.

While I appreciate and respect researchers' enthusiasm for their more recently developed methods, I suggest a healthy skepticism be employed with respect to claims of solving all problems. Every method has its shortcomings, and it is important to be aware of these before the method is used.

3 Data based bandwidth selection

The performance of all smoothing methods is crucially dependent on the specification of a smoothing parameter. This choice is a deep issue. While many things are known about "optimal bandwidth selection", they should be kept in proper perspective. There are two important issues to keep in mind. One is that in some situations there is not a single choice which gives the data analyst all the information available in the data. The other is that the classical mathematical definitions of "best" can be quite different from "useful for data analysis". See Marron and Tsybakov (1995) for reasons behind this, and some approaches to addressing this problem..

For these reasons the most useful method of choosing the amount of smoothing is interactive trial and error by an experienced analyst. But this has three important weaknesses.

- It is time consuming. Many statisticians work under time pressure, and are unable to carefully choose a smoothing parameter for every data set analyzed. Also manual choice of smoothing parameter can become very tedious for some more complicated data sets, such as the income data in Park and Marron (1990)

- Interactive choice of smoothing parameter requires some level of expertise. An inexperienced user can miss important features through oversmoothing, and can misinterpret spurious structure as being something important. Even experts are aided by a useful starting point.

- Interactive bandwidth choice is impossible when smoothers are used iteratively. This is essential to most dimensionality reduction methods, including Projection Pursuit, Generalized Additive Modelling, Alternating Conditional Expectations, and Multivariate Additive Regression Splines.

For these reasons, the subject of data-based choice of smoothing parameters is an important one.

Because simplicity allows deepest investigation, most of the research in that area has been done in the context of kernel density estimation. The starting point has been "first generation methods", such as Rule Of Thumb methods and various versions of cross-validation. These methods have their appeal, but also some serious drawbacks. The Rule Of Thumb does not make enough use of the data, and is often very inappropriate. Cross-validation based methods are widely accepted as "too noisy", meaning they can perform satisfactorily for some realizations, but they can be either under or over smoothed, just by luck of the draw, depending on the data realization (although undersmoothing is often a more serious problem).

Hence, there has been substantial effort devoted to surmounting these difficulties, which has been very successful, resulting in "second generation methods". As noted in the review/simulation papers of Park and Turlach (1992), Cao, Cuevas and Gonzales- Manteiga (1994), and Jones, Marron and Sheather (1992, 1994), there are a number of such methods available. The evidence behind these claims is firmly rooted in:

E1 Real data examples, see in particular Sheather (1992).

E2 Simulation studies.

E3 Mathematical analysis.

The whole of this evidence shows that second generation methods of bandwidth selection in kernel density estimation are ready for serious consideration by software developers.

An aside of esoteric historic interest about bandwidth selection is given in the appendix.

4 Methodological research in statistics

The evidence demonstrating the effectiveness of second generation bandwidth selection methods is compelling because it is based on real data examples, on simulation, and on asymptotic analysis. All of these tools have their strengths, and also their weaknesses. Because no method tells the whole story (and in more challenging situations the amount of the full story that is told by any can be woefully small), sensible research should involve an effective combination of as many of these as possible, especially as more challenging problems are encountered.

While this point may seem obvious, not everyone agrees. I have heard extreme viewpoints expressed on both sides of an "asymptotics vs. computation gap". If statistics becomes polarized in this way, the entire discipline will be hobbled. Strength in the field will come from proper recognition of, and respect for, work of a wide variety of types.

5 Importance of smoothing in statistics

From speaking with many statisticians, I have found a significant number who could benefit from the use of smoothing methods, but are not doing so now. I am concerned that debates about choice of smoothing method could obscure the main point that "smoothing is a useful method of analyzing data", as has been made clear for example in many books, including Cleveland (1985, 1993), Eubank (1988), Green and Silverman (1994), Härdle (1991), Hastie and Tibshirani (1990), Müller (1988), Scott (1992), Silverman (1986) and Wahba (1991).

If smoothing insiders argue too vehemently among themselves over minor technical matters, then it is easy for outsiders to lose sight of the major issues. Worse, some will develop false conceptions such as "if the experts argue so bitterly, then it must not be known how to use these methods effectively". In my opinion, the field of robustness has suffered from bickering over minor issues, thereby creating such misconceptions among the broad statistical community. I suspect this is why robust methods have been slow to gain acceptance in the wider statistical community, as exemplified by their rather sparse appearance in software packages.

Again, the key idea here is that smoothing is a useful method of analyzing data, and there are a number of effective ways to do it. While arguments over exactly how to do it become rather emotional at times, it is important that such debates remain in proper context.

6 Conclusion

Here are the main lessons that I hope can be carried away from this article:

1. There are many choices of possible smoothing methods, which are not comparable, because of differing relative strengths and weaknesses. Choice of smoother will always be a personal one, based on personal weightings of these.
2. Real progress has been made in data based bandwidth selection.
3. A combination of mathematical and computational tools provides a powerful approach to statistical problems.
4. Smoothing is an effective method to analyze data.

7 Acknowledgement

This paper was supported by NSF Grant DMS-9203135. N. I. Fisher, R. Kohn and M.P. Wand provided helpful comments.

References

[1] Cao, R., Cuevas, A. and González-Manteiga, W. (1994). A comparative study of several smoothing methods in density estimation. *Comp. Statist. Data Anal.* **17** 153-176.

[2] Cleveland, W. S. (1985) *The elements of graphing data*, Wadsworth, Belmont, California.

[3] Cleveland, W. S. (1993) *Visualizing data*, Hobart Press, Summit, New Jersey.

[4] Eubank, R. L. (1988) *Spline Smoothing and Nonparametric Regression*, Dekker, New York.

[5] Green, P. and Silverman, B. W. (1994) *Nonparametric Regression and generalized linear models, a roughness penalty approach*, Chapman and Hall, London.

[6] Härdle, W. (1991) *Applied Nonparametric Regression*, Cambridge University Press, Boston.

[7] Hall, P. (1980) Objective methods for the estimation of window size in the nonparametric estimation of a density. Unpublished manuscript.

[8] Hall, P. and Marron, J. S. (1991b) Lower bounds for bandwidth selection in density estimation. *Probab. Theory Related Fields* **90**, 149-173.

[9] Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized additive models*, Chapman and Hall, London.

[10] Jones, M. C., Marron, J. S. and Sheather, S. J. (1992) Progress in data based bandwidth selection for kernel density estimation, submitted to *J. Nonpar. Statist.*

[11] Jones, M. C., Marron, J. S. and Sheather, S. J. (1994) A brief survey of modern bandwidth selection methods, submitted to *J. Amer. Statist. Assoc.*

[12] Marron, J.S., and Tsybakov, A.B. (1995), "Visual Error Criteria for Qualitative Smoothing," *Journal of the American Statistical Association*, 90, 499-507.

[13] Marron, J. S. and Wand, M. P. (1992). Exact mean integrated squared error, *Ann. Statist.* **20**, 712-736.

[14] Müller, H.-G. (1988) Nonparametric regression analysis of longitudinal data. Springer Verlag, Berlin.

[15] Park, B.-U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *J. Amer. Statist. Assoc.* **85** 66-72.

[16] Park, B.-U. and Turlach, B. A. (1992) Practical performance of several data driven bandwidth selectors (with discussion). *Comput. Statist.* **7** 251-285.

[17] Scott, D. W. and Factor, L. E. (1981) Monte Carlo study of three data-based nonparametric probability density estimators. *J. Amer. Statist. Assoc.* **76** 9-15.

[18] Scott, D. W. and Terrell, G. R. (1987) Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.* **82** 1131-1146.

[19] Scott, D. W. (1992) Multivariate density estimation: theory, practice and visualization. Wiley, New York.

[20] Sheather, S. J. (1986). A data-based algorithm for choosing the window width when estimating the density at a point. *Comput. Statist. Data Anal.* **1** 229-238.

[21] Sheather, S. J. (1986). An improved data-based algorithm for choosing the window width when estimating the density at a point. *Comput. Statist. Data Anal.* **4** 61-65.

[22] Sheather, S. J. (1992). The performance of six popular bandwidth selection methods on some real data sets (with discussion). *Comput. Statist.* **7**, 225-250.

[23] Silverman, B. W. (1986) Density estimation for statistics and data analysis, Chapman and Hall, New York.

[24] Wahba, G. (1990) *Spline Models for Observational Data*, SIAM, Philadelphia.

8 Appendix: An historical aside

A point of esoteric historical interest is that the field of data based bandwidth selection developed rather slowly for a number of years during the mid 80's because of the following asymptotic reasoning. It is well known that, when smoothers are behaving efficiently (e.g. with respect to some norm), their asymptotic rate of convergence is determined by their "smoothness". Exactly how to quantify this is a fascinating issue (which I shall address elsewhere), but there do exist crude asymptotic surrogates, such as "number of derivatives" and "rate of tail decay of Fourier Transform". An interesting offshoot of this theory is quantification of the idea that the data contain less information about the derivative of a function, than about the function itself. This has serious implications already for "modern bandwidth selection" methods, which depend on trying to estimate derivatives. However, there are deeper conceptual problems. The current asymptotic analysis of modern bandwidth selection methods assumes more "smoothness" than the underlying estimator is capable of using, so some is "left over", and the resulting estimator will be asymptotically inefficient (compared to e.g. a "higher order kernel" method, if one is willing to ignore the bandwidth problem there). This idea held back researchers for quite some time, with some exceptions, including Hall (1980), Scott and Factor (1981) and Sheather (1983, 1985). But the ice was really broken by Scott and Terrell (1987) who showed one could get practical improvements using derivative estimation based methods, over Cross-Validation. That paper opened a flood gate of improvements by others (many of which work much better than Scott and Terrell's BCV), see Jones, Marron and Sheather (1992, 1994) for full details, discussion and references. How can the proven good performance of these methods be reconciled with the above philosophical problems? This is a deep and interesting problem. Here are three tentative ideas on this topic.

The first, as noted in Park and Marron (1990) and Hall and Marron (1991) returns to how one "attaches weights when choosing a smoother", as discussed in Section 2 above. Those who are willing to place more weight on (F2) may not be concerned about squeezing out the last epsilon of efficiency (and the results of Marron and Wand (1992) show this is pretty small all too often) from the estimator. Instead they may be quite satisfied with understanding more clearly what the estimator is telling them about the data, and with the fact that there is a very effective bandwidth choice available.

The second is that our understanding of "how asymptotics kick in" is still quite limited. Marron and Wand (1992) show that the "faster rates of convergence" associated with "more efficient estimators" can take a long time to kick in. Deeper investigation is needed, but there are indications that "what makes bandwidth selection work" kicks in at much smaller sample sizes. An interesting open problem is to develop deeper insights on this issue.

The third is that there is still not complete understanding of the abstract idea of "smoothness". To see that "number of derivatives" is only a clumsy surrogate, think about target functions which have 3.6 but not 3.7 derivatives. If I have a sample of $n = 100$ (or any number), how different are such target functions from those which have 3.7 but not 3.8 derivatives? Clearly what is needed is a more relevant way of measuring "smoothness" in finite samples. Work is under way on this problem, although it is too early to say what the outcome will be. But such ideas may be able to explain why second generation bandwidth selectors work well when current asymptotic folklore suggests otherwise.