Three Papers by Peter Bickel on Nonparametric Curve Estimation

Hans-Georg Müller[1]

ABSTRACT

The following is a brief review of three landmark papers of Peter Bickel on theoretical and methodological aspects of nonparametric density and regression estimation and the related topic of goodness-of-fit testing, including a class of semiparametric goodness-of-fit tests. We consider the context of these papers, their contribution and their impact. Bickel's first work on density estimation was carried out when this area was still in its infancy and proved to be highly influential for the subsequent wide-spread development of density and curve estimation and goodness-of-fit testing.

Key words: Asymptotics, Bias Correction, Density Estimation, Estimation of Functionals, Global deviation, Goodness-of-Fit, Nonparametric Regression.

1. Introduction

The first of Peter Bickel's contributions to kernel density estimation was published in 1973, nearly 40 years ago, when the field of nonparametric curve estimation was still in its infancy and was poised for the subsequent rapid expansion, which occurred later in the 1970s and 1980s. Bickel's work opened fundamental new perspectives, that were not fully developed until much later. Kernel density estimation was formalized in Rosenblatt (1956) and then developed further in Parzen (1962), where bias expansions and other basic techniques for the analysis of these nonparametric estimators were showcased.

Expanding upon an older literature on spectral density estimation, this work set the stage for substantial developments in nonparametric curve estimation that began in the later 1960s. This earlier literature on curve estimation is nicely surveyed in Rosenblatt (1971) and it

[1]Department of Statistics, University of California, Davis, One Shields Ave., Davis, CA 95616. Supported in part by NSF grant DMS-1104426.

defined the state of the field when Peter Bickel made the first highly influential contribution to nonparametric curve estimation in Bickel and Rosenblatt (1973). This work not only connected for the first time kernel density estimation with goodness-of-fit testing, but also did so in a mathematically elegant way.

A deep study of the connection between smoothness and rates of convergence and improved estimators of functionals of densities, corresponding to integrals of squared derivatives, is the hallmark of Bickel and Ritov (1988). Estimation of these density functionals has applications in determining the asymptotic variance of nonparametric location statistics. Functional of this type also appear as a factor in the asymptotic leading bias squared term for the mean integrated squared error. Thus the estimation of these functional has applications for the important problem of bandwidth choice for nonparametric kernel density estimates.

In the third article covered in this brief review, Bickel and Li (2007) introduce a new perspective to the well-known curse of dimensionality that affects any form of smoothing and nonparametric function estimation in high dimension: It is shown that for local linear smoothers in a nonparametric regression setting where the predictors at least locally lie on an unknown manifold, the curse of dimensionality effectively is not driven by the ostensible dimensionality of the predictors but rather by the dimensionality of the predictors, which might be much lower. In the case of relatively low-dimensional underlying manifolds, the good news is that the curse would then not be as severe as it initially appears, and one may obtain unexpectedly fast rates of convergence.

The first two papers that are briefly discussed here create a bridge between density estimation and goodness-of-fit. The goodness-of-fit aspect is central to Bickel and Rosenblatt (1973), while a fundamental transition phenomenon and improved estimation of density functionals are key aspects of Bickel and Ritov (1988). Both papers had a major impact in the field of nonparametric curve estimation. The third paper (Bickel and Li, 2007) creates a fresh outlook on nonparametric regression and will continue to inspire new approaches. Some remarks on Bickel and Rosenblatt (1973) can be found in section 2, on Bickel and Ritov (1988) section 3, and on Bickel and Li (2007) in section 4.

2. Density estimation and goodness-of-fit

Nonparametric curve estimation originated in spectral density estimation, where it had been long known that smoothing was mandatory to improve the properties of such estimates (Einstein, 1914; Daniell, 1946). The smoothing field expanded to become a major field in nonparametric statistics around the time the paper Bickel and Rosenblatt (1973) appeared. At that time, kernel density estimation and other basic nonparametric estimators of density functions such as orthogonal least squares (Čencov, 1962) were established. While many results were available in 1973 about local properties of these estimates, there had been no in-depth investigation yet of their global behavior.

This is where Bickel's influential contribution came in. Starting with the Rosenblatt-Parzen kernel density estimator

$$f_n(x) = \frac{1}{nb(n)} \sum_{i=1}^{n} w\left(\frac{x - X_i}{b(n)}\right) = \int \frac{1}{b(n)} w\left(\frac{x-u}{b(n)}\right) dF_n(u), \tag{1}$$

where $b(n)$ is a sequence of bandwidths that converges to 0, but not too fast, w a kernel function and dF_n stands for the empirical measure, Bickel and Rosenblatt (1973) consider the functionals

$$D_1 = \sup_{a_1 \le x \le a_2} |f_n(x) - f(x)|/(f(x))^{1/2}, \tag{2}$$

$$D_2 = \int_{a_1}^{a_2} \frac{[f_n(x) - f(x)]^2}{f(x)}. \tag{3}$$

The asymptotic behavior of these two functionals proves to be quite different. Functional D_1 corresponds to a maximal deviation on the interval, while functional D_2 is an integral and can be interpreted as a weighted integrated absolute deviation. While D_2, properly scaled, converges to a Gaussian limit, D_1 converges to an extreme value distribution. Harnessing the maximal deviation embodied in D_1 was the first serious attempt to obtain global inference in nonparametric density estimation. As Bickel and Rosenblatt (1973) state, *the statistical interest in this functional is twofold, as (i) a convenient way of getting a confidence band for f. (ii) A test statistic for the hypothesis $H_0 : f = f_0$.* They thereby introduce the goodness-of-fit theme, that constitutes one major motivation for density estimation and has spawned much research to this day. Motivation (i) leads to Theorem 3.1, and (ii) to Theorem 3.2 in Bickel and Rosenblatt (1973).

In their proofs, Bickel and Rosenblatt (1973) use a strong embedding technique, which was quite recent at the time. Theorem 3.1 is a remarkable achievement. If one employs a rectangular kernel function $w = 1_{[-\frac{1}{2},\frac{1}{2}]}$ and a bandwidth sequence $b(n) = n^{-\delta}$, $0 < \delta < \frac{1}{2}$, then the result in Theorem 3.1 is for centered processes

$$P\left[(2\delta \log n)^{1/2}\left([nb(n)f^{-1}(t)]^{1/2} \sup_{a_1,a_2}[f_n(t) - E(f_n(t))] - d_n\right) < x\right] \to \exp(-2\exp(-x)),$$

where

$$d_n = \rho_n - \frac{1}{2}\rho_n^{-1}[\log(\pi + \delta) + \log\log n], \quad \rho_n = (2\delta \log n)^{1/2}.$$

The slow convergence to the limit that is indicated by the rate $(\log n)^{1/2}$ is typical for maximal deviation results in curve estimation, of which Theorem 3.1 is the first. A multivariate version of this result appeared in Rosenblatt (1976).

A practical problem that has been discussed by many authors in the 1980s and 1990s has been how to handle the bias for the construction of confidence intervals and density-estimation based inference in general. This is a difficult problem. It is also related to the question how one should choose bandwidths when constructing confidence intervals, even pointwise rather than global ones, in relation to choosing the bandwidth for the original curve estimate for which the confidence region is desired (Hall, 1992; Müller et al., 1987). For instance, undersmoothing has been advocated and also other specifically designed bias corrections. This is of special relevance when the maximal deviation is to be constructed over intervals that include endpoints of the density, where bias is a particularly notorious problem.

For inference and goodness-of-fit testing, Bickel and Rosenblatt (1973), based on the deviation D_2 as in (3), propose the test statistic

$$T_n = \int [f_n(x) - E(f_n(x))]^2 a(x)\, dx$$

with a weight function a for testing the hypothesis H_0. Compared to classical goodness-of-fit tests, this test is shown to be better than the χ^2 test and incorporates nuisance parameters as needed. This Bickel-Rosenblatt test has encountered much interest; an example is an application for testing independence (Rosenblatt, 1975).

Recent extensions and results under weaker conditions include extensions to the case of an error density for stationary linear autoregressive processes that were developed in Lee and Na

(2002); Bachmann and Dette (2005), and for GARCH processes in Koul and Mimoto (2010). A related L^1-distance based goodnes-of-fit test was proposed in Cao and Lugosi (2005), while a very general class of semiparametric tests targeting composite hypotheses was introduced in Bickel et al. (2006).

3. Estimating functionals of a density

Kernel density estimators (1) require specification of a kernel function w and of a bandwidth or smoothing parameter $b = b(n)$. If one uses a kernel function that is a symmetric density, this selection can be made based on the asymptotically leading term of mean integrated squared error (MISE),

$$\frac{1}{4}b(n)^4 \int w(u)u^2\,du \int [f^{(2)}(x)]^2\,dx \;+\; [nb(n)]^{-1} \int w(u)^2\,du,$$

which leads to the asymptotically optimal bandwidth

$$b^*(n) = c\left(n\int [f^{(2)}(x)]^2\,dx\right)^{-1/5},$$

where c is a known constant. In order to determine this optimal bandwidth, one is therefore confronted with the problem of estimating integrated squared density derivatives

$$\int [f^{(k)}(x)]^2\,dx, \tag{4}$$

where cases $k > 2$ are of interest when choosing bandwidths for density estimates with higher order kernels. These have faster converging bias at the cost of increasing variance but are well known to have rates of convergence that are faster in terms of MISE, if the underlying density is sufficiently smooth and optimal bandwidths are used. Moreover, the case $k = 0$ plays a role in the asymptotic variance of rank-based estimators (Schweder, 1975).

The relevance of the problem of estimating density functionals of type (4) had been recognized by various authors, including Hall and Marron (1987), at the time the work Bickel and Ritov (1988) was published. The results of Bickel and Ritov however are not a direct continuation of the previous line of research; rather, they constitute a surprising turn of affairs. First, the problem is positioned within a more general semiparametric framework. Second, it is established that the $\sqrt{n}$ of convergence that one expects for functionals of type (4) holds if $f^{(m)}$ is Hölder continuous of order α with $m + \alpha > 2k + \frac{1}{4}$, and, with an element of surprise, that it does not hold in a fairly strong sense when this condition is violated.

The upper bound for this result is demonstrated by utilizing kernel density estimates (1), employing a kernel function of order $\max(k, m-k)+1$ and then using plug-in estimators. However, straightforward plug-in estimators suffer from bias that is severe enough to prevent optimal results. Instead, Bickel and Ritov employ a clever bias correction term (that appears in their equation (2.2) after the plug-in estimator is introduced) and then proceed to split the sample into two separate parts, combining two resulting estimators.

An amazing part of the paper is the proof that an unexpected and surprising phase transition occurs at $\alpha = 1/4$. This early example for such a phase transition hinges on an ingenious construction of a sequence of measures and the Bayes risk for estimating the functional. For less smooth densities, where the transition point has not been reached, Bickel and Rosenblatt (1973) provide the optimal rate of convergence, a rate slower than $\sqrt{n}$. The arguments are connected more generally with semiparametric information bounds in the precursor paper Bickel (1982).

Bickel and Ritov (1988) is a landmark paper on estimating density functionals that inspired various subsequent works by other authors. These include further study of aspects that had been left open, such as adaptivity of the estimators (Efromovich and Low, 1996), extensions to more general density functionals with broad applications (Birgé and Massart, 1995) and the study of similar problems for other curve functionals, for example integrated second derivative estimation in nonparametric regression (Efromovich and Samarov, 2000).

4. Curse of dimensionality for nonparametric regression on manifolds

It has been well known since Stone (1980) that all nonparametric curve estimation methods, including nonparametric regression and density estimation, suffer severely in terms of rates of convergence in high-dimensional or even moderately dimensioned situations. This is born out in statistical practice, where unrestricted nonparametric curve estimation is known to make little sense if moderately sized data have predictors with dimensions say $D \geq 4$. Assuming the function to be estimated is in a Sobolev space of smoothness p, optimal rates of convergence of Mean Squared Error and similar measures are $n^{-2p/(2p+D)}$ for samples of size n. To circumvent the curse of dimensionality, alternatives to unrestricted nonparametric regression have been developed, ranging from additive, to single index, to additive partial linear models. Due to their inherent structural constraints, such approaches come at the cost of reduced

flexibility with the associated risk of increased bias.

The cause of the curse of dimensionality is the trade-off between bias and variance in nonparametric curve estimation. Bias control demands to consider data in a small neighborhood around the target predictor levels $\mathbf{x}$, where the curve estimate is desired, while variance control requires large neighborhoods containing many predictor-response pairs. For increasing dimensions, the predictor locations become increasingly sparse, with larger average distances between predictor locations, moving the variance-bias trade-off and resulting rate of convergence in an unfavorable direction.

Using an example where $p = 2$ and the local linear regression method, Bickel and Li (2007) analyze what happens if the predictors are in fact not only located on a compact subset of $\mathcal{R}^D$, where D is potentially large, but in fact are, at least locally around $\mathbf{x}$, located on a lower-dimensional manifold with intrinsic dimension $d < D$. They derive that in this situation, one obtains the better rate $n^{-2p/(2p+d)}$, where the manifold is assumed to satisfy some local regularity conditions, but otherwise is unknown. This can lead to dramatic gains in rates of convergence, especially if $d = 1, 2$ while D is large.

This nice result can be interpreted as a consequence of the denser packing of the predictors on the lower-dimensional manifold with smaller average distances as compared to the average distances one would expect for the ostensible dimension D of the space, when the respective densities are not degenerate. A key feature is that knowledge of the manifold is not needed to take advantage of its presence. The data do not even have to be located precisely on the manifold, as long as their deviation from the manifold becomes small asymptotically. Bickel and Li (2007) also provide thoughtful approaches to bandwidth choices for this situation and for determining the intrinsic dimension of the unknown manifold, and thus the rate of effective convergence that is determined by d.

This approach likely will play an important role in the ongoing intensive quest for flexible yet fast converging dimension reduction and regression models. Methods for variable selection, dimension reduction and for handling collinearity among predictors, as well as extensions to "large p, small n" situations are in high demand. The idea of exploiting underlying manifold structure in the predictor space for these purposes is powerful, as has been recently demonstrated in Mukherjee et al. (2010) and Aswani et al. (2011). These promising approaches define a new line of research for high-dimensional regression modeling.

References

Aswani, A., Bickel, P. and Tomlin, C. (2011). Regression on manifolds: Estimation of the exterior derivative. *Annals of Statistics* **39** 48–81.

Bachmann, D. and Dette, H. (2005). A note on the Bickel-Rosenblatt test in autoregressive time series. *Statistics & Probability Letters* **74** 221–234.

Bickel, P. (1982). On adaptive estimation. *Annals of Statistics* **10** 647–671.

Bickel, P. and Li, B. (2007). Local polynomial regression on unknown manifolds. *Complex Datasets And Inverse Problems: Tomography, Networks And Beyond, ser. IMS Lecture Notes-Monograph Series.* **54** 177–186.

Bickel, P. and Ritov, Y. (1988). Estimating integrated squared densiuty derivatives: Sharp best order of convergence estimates. *Sankhya: The Indian Journal of Statistics, Series A* **50** 381–393.

Bickel, P., Ritov, Y. and Stoker, T. (2006). Tailor-made tests for goodness of fit to semiparametric hypotheses. *Annals of Statistics* **34** 721–741.

Bickel, P. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Annals of Statistics* **1** 1071–1095.

Birgé, L. and Massart, P. (1995). Estimation of integral functionals of a density. *Annals of Statistics* **23** 11–29.

Cao, R. and Lugosi, G. (2005). Goodness-of-fit tests based on kernel density estimator. *Scandinavian Journal of Statistics* **32** 599–616.

Daniell, P. (1946). Discussion of paper by M.S. Bartlett. *J. Roy. Statist. Soc. Suppl.* **8** 88–90.

Efromovich, S. and Low, M. (1996). On Bickel and Ritov's conjecture about adaptive estimation of the integral of the square of density derivative. *Annals of Statistics* **24** 682–686.

Efromovich, S. and Samarov, A. (2000). Adaptive estimation of the integral of squared regression derivatives. *Scandinavian Journal of Statistics* **27** 335–351.

Einstein, A. (1914). Méthode pour la détermination de valeurs statistiques d'observations concernant des grandeurs soumises à des fluctuations irrégulières. *Arch. Sci. Phys. et Nat. Ser. 4* **37** 254–256.

Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Annals of Statistics* **20** 675–694.

Hall, P. and Marron, J. (1987). Estimation of integrated squared density derivatives. *Statistics & Probability Letters* **6** 109–115.

Koul, H. and Mimoto, N. (2010). A goodness-of-fit test for garch innovation density. *Metrika* **71**.

Lee, S. and Na, S. (2002). On the Bickel-Rosenblatt test for first order autoregressive models. *Statistics & Probability Letters* **56** 23–35.

Mukherjee, S., Wu, Q. and Zhou, D. (2010). Learning gradients on manifolds. *Bernoulli* **16** 181–207.

Müller, H.-G., Stadtmüller, U. and Schmitt, T. (1987). Bandwidth choice and confidence intervals for derivatives of noisy data. *Biometrika* **74** 743–749.

Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics* **33** 1065–1076.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics* **27** 832–837.

Rosenblatt, M. (1971). Curve estimates. *Annals of Statistics* **42** 1815–1842.

Rosenblatt, M. (1975). A quadratic measure of deviation of two-dimensional density estimates and a test of independence. *Annals of Statistics* **3** 1–14.

Rosenblatt, M. (1976). On the maximal deviation of k-dimensional density estimates. *Annals of Probability* **4** 1009–1015.

Schweder, T. (1975). Window estimation of the asymptotic variance of rank estimators of location. *Scandinavian Journal of Statistics* **2** 113–126.

Stone, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Annals of Statistics* **10** 1040–1053.

Čencov, N. (1962). Evaluation of an unknown density from observations. *Soviet Mathematics* **3** 1559–1562.