行政院國家科學委員會專題研究計畫成果報告

區域線性迴歸平滑量由誤差決定之方法

計畫編號：NSC 88-2113-M-002-009

執行期限：87 年 8 月 1 日至 88 年 9 月 30 日

主持人:：鄭明燕副教授　國立臺灣大學數學系

一、中文摘要

我們介紹一種平滑量由誤差決定的區域線性迴歸方法。其中每個數據點的帶寬值與其誤差項絕對值的某一次方成正比，而最佳者為 2/3 次方。如此漸近變異數在常態誤差下可減少 24%，在對稱指數誤差下可減少 35%。這些結果也許顯得與 Jianqing Fan 所給的區域線性方法表現之下限矛盾，但是我們的平滑方法得到的是非線性估計量。在常態誤差下，我們的估計量的漸近均方差比 Fan 所給的對任估計量計算的極大極小下限略小。但是這些進步只真對單一函數而不是 Fan 的函數集中所有函數。當誤差項機率分佈對稱時，我們的方法對估計偏差決定項並無影響。當誤差項機率分佈不對稱時，如果適當平衡偏差與變異數我們的方法仍可減低估計均方差。

Abstract:

An error-dependent smoothing rule for reducing the variance of local linear curve estimators is suggested. It involves weighting the bandwidth used at each datum, in proportion to a power of the absolute value of its residual. The optimal power is 2/3. We prove that asymptotic variance can be reduced by 24% in the case of Normal errors, and 35% for double-exponential errors. These results might appear to violate Jianqing Fan's bounds on performance of local linear methods, but note that our approach to smoothing produces nonlinear estimators. Under Normal errors, our estimator has slightly better mean squared error performance than that suggested by Fan's minimax bound, calculated over all estimators. However, these improvements are available only for single functions, not uniformly over Fan's function class. For symmetric error distributions the method has no first-order effect on bias. In the case of asymmetric error distributions an overall reduction in mean squared error is achievable, involving a trade-off between bias and variance contributions.

Keywords: Bandwidth, kernel method, nonparametric regression, tail weight, variance reduction.

二、緣由與目的

There is a great variety of bias reduction methods for nonparametric curve estimation, ranging from high-order kernel techniques (e.g. Wand and Jones, 1995, Chapters 2 & 5) to local bandwidth

adjustments (e.g. Abramson, 1982), methods based on varying location and scale (e.g. Samiuddin and el-Sayyad, 1990) empirical transformations (e.g. Ruppert and Cline, 1994), weights (e.g. Jones, Linton and Nielsen, 1995), and skew computation (e.g. Choi and Hall, 1998). However, very few methods have been suggested for reducing the impact of variance. Those that do exist involve principally deterministic adjustments to bandwidth, altering the local trade-off between variance and squared bias in the context of mean integrated squared error. We suggest a new and entirely different approach to variance reduction. It involves adjusting bandwidth in a stochastic way, with the aim of providing improved performance by giving greater weight to data pairs that correspond to smaller absolute errors. The method is also applicable to the case of heteroscedasticity.

三、結果與討論

Assume that $(X_1,Y_1),\cdots,(X_n,Y_n)$ are independent and identically distributed data pairs generated by the model

$$Y_i = g(X_i)+\varepsilon_i, \qquad (1)$$

where X_i is independent of ε_i, and $E(\varepsilon_i)=0$. In the ``ideal" case, where the errors ε_i are known, define a bandwidth h_i by $h_i = hH(\varepsilon_i)$, with $h=h(n)$ being a sequence of positive constants and H a fixed positive function. Realistically, we approximate ε_i by a residual $\hat{\varepsilon}_i$ and put $\hat{h}_i = hH(\hat{\varepsilon}_i)$. In either case, and for fixed x, let $(\hat{a},\hat{b})$ denote the pair that minimizes

$$\sum_{i=1}^{n}\{Y_i - a - b(X_i - x)\}^2 h_i^{-1} K\{(X_i - x)/h_i\},$$

where K is a kernel function; and put $\hat{g}(x)=\hat{a}$. Here K is taken to be a bounded, compactly supported, symmetric probability density. We expect $H(x)$ to be an increasing function of $|x|$ in which case smaller absolute errors produce lesser smoothing. Suppose H is non-degenerate. If $E\{H(\varepsilon)^2\}<\infty$ then we may standardize H so that $E\{H(\varepsilon)^2\}=1$. In addition, provided the error distribution is symmetric and H is an even function, then

$$\hat{g} = g + \frac{1}{2}h^2 g''\kappa_2 + (nh)^{-\frac{1}{2}}\rho f^{-\frac{1}{2}}\kappa^{\frac{1}{2}}\sigma N_n + o_p(h^2), \qquad (2)$$

with $\kappa_2 = \int u^2 K(u)du, \kappa = \int K^2, \sigma^2 = \mathrm{var}(\varepsilon)$, $\rho^2 = E\{\varepsilon^2 H(\varepsilon)^{-1}\}/\sigma^2 < 1$, and N_n is a random variable whose asymptotic distribution is N(0,1). Here, \$f\$ denote the marginal density of X_i, assumed to exist in a neighborhood of x and $f(x)>0$. Note that the expansion (2) holds for the classical local linear estimator, except that the second term does not contain the factor ρ. The solution to minimizing ρ subject to $E\{H(\varepsilon)^2\}=1$ is that $H(\varepsilon)$ proportional to $|\varepsilon|^{2/3}$, in which case $\rho^2=\rho_0^2$ where

$$\rho_0^2 = \{E(|\varepsilon|^{4/3})\}^{3/2}/\sigma^2 < 1.$$

For Normal errors, our results are suggestive of a version of Fan's (1993) Theorem 4 in which his constant 0.869^2 is replaced by 1.00, tis being the value (to two decimal places) of

$$(1.243\rho_0^{8/5})^{-1} = \left(1.243\left[2\{\Gamma(7/6)/\pi^{1/2}\}^{3/2}\right]^{4/5}\right)^{-1} \approx 1.0024^2. \qquad (3)$$

The fact that right-hand side of (3) exceeds 1

means that a realistic version of the nonlinear ideal estimator $\hat{g}$ cannot be expected to achieve, in a uniform sense. Therefore, Fan's (1993) results do not have close analogues for our estimator. However, for any particular functions f and g with two continuous derivatives, the level of performance evinced by the ideal estimator can be achieved. In this case a pilot estimator $\tilde{g}$ is first constructed, and used to calculate $\hat{\varepsilon}_i = Y_i - \tilde{g}(X_i)$. Then, possibly after centering or thresholding these residuals, we would compute the empirical bandwidth $\hat{h}_i = h|\hat{\varepsilon}_i|^{2/3}$ and use $\hat{h}_i$ in place of h_i to construct $\hat{g}$. Modulo regularity conditions, formula (2) continues to hold in this case.

Since Fan's minimax function class C_2 contains models where the error distribution is Normal, results such as those discussed above do not relate to improvements in performance that error-dependent smoothing can achieve in the case of heavy-tailed error distributions. Indeed, the value of ρ_0^2 tends to be smaller for distributions with heavier tails. Values in the cases of double Exponential, Normal and Uniform errors are respectively 0.650, 0.757 and 0.842. For errors with Student's t distribution on 5, 10 or 20 degrees of freedom, the values are 0.676, 0.726 and 0.743, respectively. For non-Normal errors one can also construct estimators that are more efficient than $\hat{g}$ by replacing local least-squares by a robust method such as local M-estimation.

Neither (2) nor (3) is available uniformly in a function class such as Fan's C_2. One reason is that the rate at which the "$o_p(h^2)$" terms converge to 0 depends explicitly on the modulus of continuity of both f'' and g'', and can be arbitrarily slow. In the realistic case, this in turn influences choice of the pilot-estimator bandwidth. Taking that quantity to be equal to a fixed constant multiple of $n^{-1/5}$ results in inflation of variance relative to that achieved by the "ideal" estimator. For sufficiently heavy-tailed error distributions, such an inflation still produces a reduction in variance, relative to that for a classical local linear estimator. This is readily apparent in both theoretical analysis and numerical simulations. For Normal data, however, slight oversmoothing of the pilot estimator, and relatively large sample sizes, are necessary in order to achieve obvious improvements.

In a simulation study where the error distribution was Student's t with 5 degrees of freedom, we found that when $n = 100$ the "ideal" and "realistic" error-dependent smoothing rules reduced mean integrated squared error (MISE) by 30% and 13% respectively. Greater reductions occurred for larger values of n. In the case of Normal errors, however, MISE reductions in the "realistic" case only became significant for $n = 500$. In this work we took $g(x) = 4\sin(2\pi x)$, used equally-distributed design points on $(0,1)$, employed the biweight kernel, and took the bandwidth for the pilot estimator to be 25% greater than its conventional asymptotically optimal value.

四、計劃成果自評

In large samples, while keeping the same bias as the classical local linear estimators, our

method reduces the variance contribution to mean integrated squared error by 24% in the case of Normal errors, by 35% for double-exponentially distributed errors, and by even greater amounts for very heavy-tailed error distributions. These figures might appear to violate the bounds asserted by Fan (1993) for performance of local linear estimators. However, Fan's result about asymptotic optimality of local linear estimators applies only within the class of linear techniques, and (after error-dependent smoothing) our estimators are nonlinear. Moreover, Fan's other minimax bounds, measuring performance against nonlinear techniques, do not specifically address the range of heavy-tailed error distributions that are considered in the present paper, since his class C_2 of models for the ``max'' part of ``minimax'' contains models with Normal errors.

五、參考文獻

1. Abramson, I.S. (1982). On bandwidth variation in kernel estimates --- a square root law. *Ann. Statist.* **9**, 168--176.
2. Choi, E. and Hall, P. (1998). On bias reduction in local linear smoothing. *Biometrika* **85**, 333--346.
3. Fan, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196--216.
4. Jones, M.C., Linton, O. and Nielsen, J.P. (1995). A simple bias reduction method for density estimation. *Biometrika* **82**, 327--338.
5. Ruppert, D. and Cline, B.H. (1994). Bias reduction in kernel density estimation by smoothed empirical transformations. *Ann. Statist.* **22**, 185--210.
6. Samiuddin, M. AND El-Sayyad, G.M. (1990). On nonparametric kernel density estimates. *Biometrika* **77**, 865--874.
7. Wand, M.P. and Jones, M.C. (1995). *Kernel smoothing*. Chapman and Hall, Lordon.