行政院國家科學委員會補助專題研究計畫 ☑ 成 果 報 告
□期中進度報告

估計二維函數曲面(2/2)

計畫類別：☑ 個別型計畫 □ 整合型計畫
計畫編號：NSC 90－2118－M－002－011－
執行期間： 90 年 8 月 1 日 至 91 年 10 月 31 日

計畫主持人： 鄭明燕
共同主持人：
計畫參與人員： 林思成、顏嘉宏、楊玲惠

成果報告類型(依經費核定清單規定繳交)：☑精簡報告 □完整報告

本成果報告包括以下應繳交之附件：
☑赴國外出差或研習心得報告一份
□赴大陸地區出差或研習心得報告一份
□出席國際學術會議心得報告及發表之論文各一份
□國際合作研究計畫國外研究報告書一份

處理方式：除產學合作研究計畫、提升產業技術及人才培育研究計畫、列管計畫及下列情形者外，得立即公開查詢
□涉及專利或其他智慧財產權，□一年□二年後可公開查詢

執行單位： 臺灣大學數學系

中 華 民 國 92 年 1 月 27 日

中文摘要：

我們研究以非參數數方法估計被二維機率密度函數或具兩個解釋變數之迴歸函數定義的曲面。我們建議一種有效減低變異量的估計方法。此方法可用於許多推論問題如機率密度估計及迴歸，亦可用於許多型式的估計量。首先我們利用傳統的曲面估計量建造一個估計曲面等高線的一次或二次曲線線段，建造方式乃使得傳統的曲面估計量沿此線段有最小方差。有了此等高線估計線段，最終的曲面估計值是所有沿著此線段一部份的傳統估計值的平均值。我們探討應用此方法至估計機率密度函數曲面的核估計量。我們探討這些估計量的理論性質與數值表現。

關鍵詞：帶寬值、邊界效應、核方法、無母數機率密度估計、無母數迴歸、變異降低。

Abstract:
We suggest a method for reducing variance in nonparametric surface estimation. The technique is applicable to a wide range of inferential problems, including both density estimation and regression, and to a wide variety of estimator types. It is based on estimating the contours of a surface by minimizing deviations of of elementary surface estimates along a linear or quadratic curve. Once a contour estimate has been obtained, the final surface estimate is computed by averaging conventaional surface estimates along a portion of the contour. Theoretical and numerical properties of the technique are discussed.

Keywords: Bandwidth, boundary effect, kernel method, nonparametric density estimation, nonparametric regression, variance reduction.

前言：We study nonparametric estimation of surfaces defined by a univariate function of a 2-vector. Estimation of a bivariate surface is much more challenging compared to estimation of one-dimensional curves. For example, scales of the two separate directions can be substantially different and the smoothing parameters need to be carefully chosen and boundary curve of the support can be complex. Moreover, data sparseness problems seriously affect the performance.

研究目的：Let g denote a univariate function of a 2-vector; for example, g might be the density of a bivariate distribution, or the mean in a regression problem where the explanatory variable is bivariate and the response variable is scalar. We wish to estimate g nonparametrically, making only smoothness assumptions and exploiting the extra degree of freedom that is available in the context of surface estimation.

文獻探討：Scott (1992, p.~149ff) and Hall, Huber, Owen and Coventry (1994) discuss the so-called balloon kernel techniques for density estimation. There is an extensive literature on approaches for remedying boundary effects in density estimation and regression, mainly in univariate cases. It includes methods based on special ``boundary kernels," for example those considered by Gasser and Mueller (1979), Gasser, Mueller and Mammitzsch (1985), Granovsky and Mueller (1991) and Mueller (1991). Rice (1984) suggested a dual-bandwidth approach. So-called ``reflection methods'' include those of Schuster (1985), and Hall and Wehrly (1991) Cheng, Fan and Marron (1997) suggested methods that have optimal asymptotic performance at boundaries. The natural boundary-respecting properties of local polynomial methods have been discussed by Fan (1993), Ruppert and Wand (1994), and Fan and Gijbels (1996), for example.

研究方法：Let $¡ = \{X_1,..., X_n\}$ be a sample observed from a bivariate density function g. We first construct a nonparametric kernel estimator $\hat{g}$ of g. Let $C(x\,|\,¯, c)$ be the parabola with its vertex at $x = (x^{(1)}, x^{(2)})$ and its tangent there in the direction of the unit vector $(\cos ¯, \sin ¯)$ and with curvature $2c$ at x. We constrain $¯$ and c by $-®/2 < ¯ \le ®/2$ and $-\infty < c < \infty$ which ensures that each nondegenerate parabola $C(x\,|\,¯, c)$ in the plane is represented by for a unique triple $(x, ¯, c)$. Given $ª > 0$, let $C(x\,|\,¯, c, ª)$ denote the set of points $z \in C(x\,|\,¯, c)$ that satisfy $\|z - x\| \le ªh$. Put $¶(c, ª) = |C(x\,|\,¯, c.\,ª)|$,

$$\breve{g}(x\,|\,¯, c, ª) = ¶(c, ª)^{-1} \int_{C(x|¯,c,ª)} \hat{g}(z)\,ds,$$

$$S(x\,|\,¯, c, ª) = ¶(c, ª)^{-1} \int_{C(x|¯,c,ª)} \{\hat{g}(z) - \breve{g}(x\,|\,¯, c, ª)\}^2\,ds,$$

$$(\hat{¯}_x, \hat{c}_x) = \arg\min_{(¯,c)} S(x\,|\,¯, c, ª),$$

where ds is an infinitesimal element of arc length along $C(x\,|\,¯, c, ª)$. Our final estimator of g is $\tilde{g}(x\,|\,ª) = \breve{g}(x\,|\,\hat{¯}_x, \hat{c}_x, ª)$. A variant is to replace the parabola segment $C(x\,|\,¯, c, ª)$ by a line segment, which results in a linear approximation.

結果與討論：A relatively large value of $ª$ (asymptotically, $ª \to \infty$) should be used to give accurate estimation of the true quadratic approximation to the contour line at x. On the other hand, a relatively small value of $ª$ may be adquate for reducing variance and removing edge effects.

We show that the estimator is a uniformly good approximation to its ideal but impractical counterpart obtained by averaging the conventional kernel estimator along a portion of the true contour line passing through x. Our asymptotic results imply that the asymptotic variance of our estimator is strictly less than that of the conventional kernel density estimator. The reduction of asymptotic variance is by a constant factor.

There is a natural version of our estimator that address edge effects. It is straightforward to state and derive versions of the asymptotic results in this context.

In a simulation study, it was observed that our method produces much less wiggle estimates than the kernel density estimate. Furthermore, by a simple choice of $ª$ it provides satisfactory estimates of various types, for example unimodal and bimodal, of density surfaces.

計劃成果自評： Our proposed method has several novel features. First, it exploits the extra degree of freedom that is available in the problem of surface estimation. Secondly, it provides a new technique for estimating gradients and curvatures of contour lines, without passing explicitly to derivatives of surface estimates. Thirdly, when applied to a surface estimate that is always positive, in either density estimation or regression, our method produces a boundary-corrected estimate that is always positive. The technique is applicable to nonparametric methods in both density estimation and regression. Indeed, it is not tied to a particular estimator type in either of these settings; for example, in nonparametric regression it can be used in conjunction with spline, local linear or Nadaraya-Watson methods. In the case of density estimation, when a conventional kernel estimator is used as its basis, the technique can be viewed as a device for re-computing kernel shape. As a result of this research grant, we have produced a paper ``Reducing variance in nonparametric surface estimation'' (Cheng and Hall, 2003) which has been accepted by the Journal of Multivariate Analysis.

參考文獻：

1. CHENG, M.-Y., FAN, J. AND MARRON (1997) On automatic boundary corrections. *Ann. Statist.*, **25**, 1691—1708.
2. CHENG, M.-Y. AND HALL, P. (2003) Reducing variance in nonparametric surface estimation, *Journal of Multivariate Analysis*, to appear.
3. FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21**, 196--216.
4. FAN, J. AND GIJBELS, I. (1996). *Local Polynomial Modelling and its Applications*. Chapman and Hall, London.
5. GASSER, T. AND MUELLR, H.G. (1979). Kernel estimation ofbregression functions. *Smoothing Techniques for Curve Estimation.* Lecture Notes in Mathematics, **757**, 23--68. Springer, New York.
6. GASSER, T., MUELLER, H.G. AND MAMMITZSCH, V. (1985). Kernels for nonparametric curve estimation. J*. Roy. Statist. Soc. Ser. B,* **47**, 238--252.
7. GRANOVSKY, B.L AND MUELLER, H.G. (1991). Optimizing kernel methods: a unifying variational principle. *Internat. Statist. Rev.* **59**, 373--388.
8. HALL, P. AND WEHRLY, T.E. (1991). A geometrical method for removing edge effects from kernel-type nonparametric regression estimators. *Amer. Statist. Assoc.,* **86**, 665--672.
9. HALL, P., HUBER, C., OWEN, A. AND COVENTRY, A. (1994). Asymptotically optimal balloon density estimates. *J. Multivariate Anal.* **51**, 352--371.
10. LEJEUNE, M. AND SARDA, P. (1992).Smooth estimators of distribution and density functions. *Comput. Statist. Data Anal.*, **14**, 457—471.
11. MUELLER, H.G. (1991). Smooth optimal kernel estimators near endpoints. *Biometrika* **78**, 521--530.
12. RICE. J. (1984). Boundary modification for kernel estimators. *Comm. Statist. Theory Methods* **13**, 893—900.
13. RUPPERT, D. AND WAND, M.P. (1994). Multivariate locally weighted least squares

regression. *Ann. Statist.* **22**, 1346--1370.

14. SCHUSTER, E.F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Comm. Statist. Theory Methods*. **14**, 1123--1136.
15. SCOTT, D.W. (1992). *Multivariate Density Estimation -- Theory, Practice and Visualization.* Wiley, New York.