行政院國家科學委員會專題研究計畫　成果報告

多維及相關數據的非參數估計問題(2/2)

計畫類別： 個別型計畫
計畫編號： NSC92-2118-M-002-012-
執行期間： 92 年 08 月 01 日至 93 年 10 月 31 日
執行單位： 國立臺灣大學數學系暨研究所

計畫主持人： 鄭明燕

計畫參與人員： 林思成、楊幼君、蕭光呈、沈愈騰

報告類型： 完整報告
報告附件： 國外研究心得報告
處理方式： 本計畫可公開查詢

中　華　民　國 94 年 2 月 2 日

中文摘要：在生物、經濟等領域中，多維變量的值域邊界是一個重要問題，我們提出一種計算簡易的估計方法並且證明它有最佳的漸近收斂速度。非參數曲面估計已被廣泛應用於實務，其中多維區域線性迴歸估計有許多優點，我們提出數種降低多維區域線性迴歸估計變異方法，發現皆可大幅降低漸近變異而不影響漸近誤差。

關鍵詞：多維區域線性迴歸、非參數平滑方法、值域邊界、降變異法

英文摘要：In biological and econometrical sciences, it is often important to know the support boundary of a multivariate random vector. We propose a method that is easy to compute and prove that it achieves the optimal rate of convergence. Nonparametric estimation of surfaces has been widely used in practice. Multivariate local linear regression has many advantages. We suggest methods to reduce variance of multivariate local linear smoothers. It is shown that they all substantially enhances the stability while keeping the asymptotic bias unchanged.

英文關鍵詞：multivariate local linear regression, nonparametric smoothing, support boundary, variance reduction.

前言：In a range of practical problems the boundary of the support of a bivariate distribution is of interest, for example where it describes a limit to efficiency or performance, or where it determines the physical extremities of a spatially distributed population in forestry, marine science, medicine, meteorology or geology. We suggest a tracking-based method for estimating a support boundary when it is composed of a finite number of smooth curves, meeting together at corners.

Kernel methods for nonparametric estimation of surfaces have been widely used in applications. Multivariate local linear regression enjoys many advantages. It is minimax optimal among all linear estimators, automatically achieves boundary corrections, adapts to both random and fixed designs. We study improvements of multivariate local linear regression. Two intuitively appealing variance reduction techniques are proposed. They both yield estimators that retain the same asymptotic conditional bias as the multivariate local linear estimator and have smaller asymptotic conditional variances.

文獻探討：Tracking methods were suggested by Hall and Rau (2000) for solving a fault-type regression problem, in the absence of corners. There is an extensive literature on non-tracking methods for boundary-support estimation when the curve contains no corners. A significant part of it is in the area of econometrics, where the boundaries are often interpreted as ``production frontiers".  Many nonparametric techniques are based on enveloping the data in some sense, and include ``data envelopment analysis" (Farrell, 1957) and the ``free disposal hull" (Deprins, Simar and Tulkens, 1984). Theoretical performance (including convergence rates) and numerical properties of these and other methods have been investigated by Korostelev and Tsybakov (1993), Mammen and Tsybakov (1995) and others. However, these treatments do not permit corners, and do not address tracking methods.

Nonparametric regression methods are useful for exploratory data analysis and for representing underlying features that can not be well described by parametric regression models. In the recent two decades, many attentions have been paid to local polynomial modeling for nonparametric regression which was first suggested by Stone (1977). Ruppert and Wand (1994) established theoretical results for local polynomial regression with multiple covariates.

研究方法：Assume we observe data from a realization of a point process in the plane. The boundary will be traced in a clockwise direction, and so ``below" may equivalently be thought of as lying to the right of the direction of travel, although we shall use ``left" and ``right" for another purpose. The notion of a short line segment that has no points above it is intuitively clear in many cases. Our boundary estimator is piecewise linear, and in particular consists of line segments joining adjacent estimators Q_j of points on the support boundary, indexed in such a manner that we move in a clockwise sense. We pass from Q_j to Q_{j+1} by first moving to a preliminary point Q_{j+1}', calculated by fitting either a left smooth or a right smooth to the boundary at Q_j; and then we refine Q_{j+1}' to Q_{j+1} by fitting two short line segments to data in the vicinity of Q_{j+1}'. This procedure by itself produces a boundary estimator that tends to cut across the corners, however, rather than reach into them. As a result, this simple form of the boundary estimator does not enjoy the desired level of accuracy. To overcome this problem we suggest a threshold technique for deciding when the sequence has cut a corner. We discard a subsequence that cuts a corner, close up the remaining members of the sequence, and estimate the corner by extrapolating to it from points Q_j that lie on either side of the discarded sequence. These operations are conducted completely sequentially, and in particular do not involve drawing the boundary and then erasing part of it. Our algorithm tells the curve estimator unambiguously when to mark time, i.e. to stop confirming boundary points $\hQ_j$, and when to start confirming them again.

Our first idea of variance reduction in multivariate local linear regression is the following. Given the local linear estimates evaluated over a grid, we form a linear combination of the estimates, to be a new estimate. In this way the resultant estimate is not allowed to differ too much from the estimates over the grid, and its source of variability is restricted to their variances and covariances. Meanwhile, to ensure that the asymptotic conditional bias unchanged, the new estimator has to be subject to certain moment conditions. This can be accomplished by forcing the coefficients in the linear combination to fulfill the corresponding requirements. Our first variance reducing estimator improves the original local linear estimator in a non-uniform manner. The best pointwise relative variance reduction occurs at the some specific points. Our second variance reducing estimator is then constructed to achieve this best relative efficiency everywhere. The approach is that, fixing at each point of estimation, evaluate the usual local linear estimates at points surrounding it and then linearly combine these estimates to form a new estimator in the same way as in the first method.

結果與討論：Our support boundary estimator achieves the uniform convergence rate in the absence of corners. The rate in the presence of corners must therefore also be optimal. With the minor alteration that our point process is Poisson, rather than the result of distributing a given number of independent random variables, we obtain an analogous result in the presence of corners .

The variance reduction estimators both yield estimators that retain the same asymptotic conditional bias as the multivariate local linear estimator and have smaller asymptotic conditional variances. The estimators are further examined in aspects of bandwidth selection, asymptotic relative efficiency and implementation. Their asymptotic relative efficiencies with respect to the multivariate local linear estimator are very attractive and increase exponentially as the number of covariates increases. Data-driven bandwidth selection procedures for the new estimators are straightforward given those for local linear regression. Since the proposed estimators each has a simple form, implementation is easy and requires much less or about the same amount of effort. In addition, boundary corrections are automatic as in the usual multivariate local linear regression.

計劃成果自評：Two papers are generated from this one-year project: Cheng and Hall (2004) “Methods for tracking support boundaries with corners ” and Cheng and Peng (2004) “Simple and efficient improvements of multivariate local linear regression.” Both have been submitted to *Journal of Multivariate Analysis*.

參考文獻：

1. DESHAYES, J. AND PICARD, D. (1981). Convergence de processusa double indice: application aux tests de rupture dans unmodele. *Comptes Rendus* **292**, 642--669.
2. FARRELL, M.J. (1957). The measurement of productive efficiency. *J. Roy. Statist. Soc. Ser. A* **120**, 253--281.
3. HALL, P. AND RAU, C. (2000). Tracking a smooth fault line in a response surface. *Ann. Statist.* **28**, 713--733.
4. KOROSTELEV, A.P. AND TSYBAKOV, A.B. (1993). Estimating the support of a density and functionals of it. *Problems Inform. Transmission* **29**, 1--15.
5. MAMMEN, E. AND TSYBAKOV, A.B. (1995). Asymptotical minimax recovery of sets with smooth boundaries. *Ann. Statist.* **23**, 502--524.
6. RUPPERT, D. AND WAND, M.P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.* **22**, 1346--1370.
7. STONE, C.J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595--645.