行政院國家科學委員會專題研究計畫 成果報告

處理微晶片數據之相關多變量分析及模型選取
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 95-2118-M-002-004-
執 行 期 間 ：95 年 08 月 01 日至 96 年 07 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：陳宏

計畫參與人員：博士班研究生-兼任助理：倪惠芬、葉倚任
碩士班研究生-兼任助理：金妍秀、黃以達、黃信雄、侯坤穗
臨時工：吳欣屏

報 告 附 件 ：出席國際會議研究心得報告及發表論文

處 理 方 式 ：本計畫可公開查詢

中 華 民 國 96 年 12 月 18 日

行政院國家科學委員會補助專題研究計畫成果報告

處理微晶片數據之相關多變量分析及模型選取

Research on Multivariate Analysis and Model Selection for Microarray Data

計畫類別：■個別型計畫　　□ 整合型計畫
計畫編號：NSC　95－2118－M－002－002－
執行期間：2006 年 8 月 1 日至 2007 年 7 月 31 日

計畫主持人：陳宏
共同主持人：
計畫參與人員：

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：國立台灣大學數學系

中　華　民　國 九十六 年　十二 月 十六 日

一、中文摘要

本計畫探討cDNA 微晶片實驗所產生的問題，第一個問題是決定一極大的ANOVA模型中的可估參數，這個模型乃用於數據的正規化；第二個問題是探討線性迴歸中使用L_1限制下之模型選取問題及探討基因選取之多重假設檢定問題。

關鍵詞：cDNA 微晶片數據，數據正規化，模型選取，多重假設檢定。

Abstract

In this project, we study two problems arose in cDNA microarray data analysis. The first problem is on identifying estimable parameters in a large two-way additive ANOVA models which is related to the normalization. The second problem is on selecting informative variables in linear regression model with large number of unknown parameters using L_1 norm constraint and multiple testing problem arose in gene selection.

Keywords: cDNA Microarray data, normalization, model selection, multiple testing.

二、Consistent Estimate of Component in a Bivariate Additive Model with Sparse Data

Motivated by “local normalization” to remove bias in the observed intensity levels of gene expressions measured by microarray study, we consider a bivariate additive model in which the intensity effect is modeled by a smooth function. When the smooth function is approximated by a regression spline, it is shown that the estimate of smooth component can no longer achieve the usual rate of convergence as that of univariate nonparametric regression.

Nonparametric regression has been used in various applications to explore the relationship between dependent variable and predictors. Due to the curse of dimensionality, general nonparametric regression maynot be useful when there are many predictors. Instead, additive regression is considered in Stone (1985) which can be used in a wide variety of situation for readily interpretable. Stone (1985) showed that the additive regression can be estimated with the usual one-dimensional convergence rate under proper design condition. Under similar design condition as in Stone (1985), Opsomer and Ruppert (1997) gave a detailed analysis on the bias of a bivariate additive regression.

As motivated by the result in Fan et al. (2005), we consider estimation problem in the following bivariate additive model

$$y = m_1(x_1) + m_2(x_2) + \epsilon, \tag{1}$$

where y is the response variable, $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ are covariates. Under the setting of Fan et al. (2005) and the print-tip block effects are known, x_1 is a categorical variable taking

values on a finite set $\mathcal{X}_1$, $\mathcal{X}_2$ is an interval, $m_1(\cdot)$ is bounded, and $m_2(\cdot)$ is a smooth function with $\mathrm{E}[m_2(x_2)] = 0$. Moreover, there are I observations associated with each element in $\mathcal{X}_1$ while the cardinality is of the same order the number of total observations, where $I \geq 2$.

This paper addresses the question on identifying design condition to ensure the asymptotic result as given in Fan et al. (2005) under (1). The objective is achieved by considering how to aggregate information from the I observations and how to express the information matrix of the parameters used to approximate $m_2(\cdot)$. Then, under the general design condition, the asymptotic result in Fan et al. (2005) is attempted to connect with which in Stone (1985) and Opsomer and Ruppert (1997).

By writing the information in terms of a mixture of covariance matrix of multinomial distributions, we not only give a new proof why "connectedness" is a necessary condition on the estimability of a two-way additive model, but also we can get the result as in Fan et al.~(2005) under much more general design condition.

三、Operating characteristics of C_p-LASSO on variable selection in linear regression with orthonormal regressors

Model selection coupled with regularization is a commonly used method to do model fitting to achieve sparsity or parsimony of resulting model. In this project, we study the operating characteristics of LASSO (Tibshirani, 1996) coupled with Mallows' C_p on identifying important orthonormal predictor variables of linear regression. We consider the case that the dimensionality of predictor variables, *m*, is high and the number of observations, *n*, is of the same order *m*.

The orthogonal predictors arises naturally in the problem of nonparametric function estimation with a wavelet basis or through the conversion of nonorthogonal predictor variables by principal component analysis. When the goal in variable selection is to select a model to minimize the mean square error of prediction, Mallows C_p (1973) is the often used penalized model selection criteria to achieve it by combining the residual sum of squares and the fitted number of predictors. Efron et al. (2004) also suggest the use of Mallows' C_p to do variable selection with Lasso.

For a given nested linear models, a common approach to the selection of statistical models is the so-called penalized model selection criteria which include Mallows' C_p (1973). Woodrofe (1982) and Zhang (1992) give a detailed description on the number of uninformative predictors chosen by Mallows' C_p when one of the nested linear model is the correct one. It is shown that Mallows' Cp leads to a over-fitted regression model with no more than one noninformative predictor in average. For a linear regression model with uncorrelated predictors, Lasso gives a data-driven nested models by varying bound on the L_p norm of the coefficients as stated in Efron, Hastie, Johnstone and Tibshirani (2004). In this paper, an analysis along the line of Woodrofe (1982) and Zhang (1992) to describe the operating characteristic of using Mallows' C_p as the automatic predictor selection criterion for the Lasso method when the number of predictors is large and all informative predictors are among $\{X_1, \ldots, X_m\}$. The reported result also addresses the comments made by Ishwaran and Stine on their discussions of the use of C_p-Lasso shrinkage in Efron et al. (2004).

We now briefly review the Lasso method in the linear regression model with uncorrelated predictors. We also characterize the random walk induced by C_p-Lasso under normal error.

Suppose that we have data $(\mathbf{x}^i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}^i = (x_{i1}, \ldots, x_{im})^T$ are the predictor variables and y_i are the responses. Let $\mathbf{X}_n$ by the $n \times m$ predictor matrix with ijth entry x_{ij} and $\mathbf{X}_n^T \mathbf{X}_n = (n/m)\mathbf{I}_m$, where $\mathbf{I}_m$ is the $m \times m$ identity matrix. We assume a homoskedastic model

$$y_i = \alpha + \sum_{j=1}^{m} \beta_j x_{ij} + \epsilon_i, \tag{1}$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d. normal random variables with mean 0 and variance σ^2. Let $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_m)^T$ denote the vector of model parameters and m_0 denote the total number of nonzero β_j.

The Lasso estimate is defined by

$$\sum_{i=1}^{n} \left(Y_i - \alpha - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 \quad \text{subject to } \textstyle\sum_{j=1}^{m} |\beta_j| \leq \tau. \tag{2}$$

Without loss of generality, we can assume $\sum_{i=1}^{n} x_{ij} = 0$ for $j = 1, 2, \ldots, m$ by employing location transformation when none of $\mathbf{x}_i$ falls in the linear space spanned by $\mathbf{1} = (1, \ldots, 1)^T \in R^n$. Then the minimization problem in (2) can be solved as the following minimization problem

$$\sum_{i=1}^{n} \left(y_i - \bar{y} - \sum_{j=1}^{m} \beta_j x_{ij} \right)^2 \quad \text{subject to } \textstyle\sum_{j=1}^{m} |\beta_j| \leq \tau. \tag{3}$$

We now describe the Lasso algorithm for accruing predictor variables one at a time in successive steps by decreasing the penalty τ accordingly. Let $\hat{\beta}_j$, $1 \leq j \leq m$, denote the least squares estimates of β_j. Lasso adds the kth predictor X_j which satisfies $\mathbf{x}_j^T \mathbf{y} = |\hat{\beta}|_{(m-k)}$, where $|\hat{\beta}|_{(1)} < \cdots < |\hat{\beta}|_{(m)}$. At the kth step, Lasso leads to

$$\hat{\beta}_{kj} = sign(\hat{\beta}_j)(|\hat{\beta}_j| - |\hat{\beta}|_{(m-k)})^+$$

and

$$\hat{\mu}_k = \bar{y}\mathbf{1} + \sum_{i=1}^{m} \left[\left(\hat{\beta}_j - |\hat{\beta}|_{(m-k)} \right) 1(\hat{\beta}_j > |\hat{\beta}|_{(m-k)}) + \left(\hat{\beta}_j + |\hat{\beta}|_{(m-k)} \right) 1(\hat{\beta}_j < -|\hat{\beta}|_{(m-k)}) \right] x_i$$

where $1(\cdot)$ is the indicator function and $(\pi)^+ = \pi$, $\pi > 0$; 0, $\pi \leq 0$.

By decreasing τ from $|\hat{\beta}|_{(m)}$ to $|\hat{\beta}|_{(1)}$, Lasso gives a data-dependent nested models. C_p-Lasso is to use Mallows' C_p to identify important predictors among the data-driven nested models. At the kth step of Lasso, Mallows' C_p is defined to be $\|\mathbf{y} - \hat{\mu}_k\|^2/\sigma^2 + (2k - n)$. ¿From now on, we assume that σ^2 is known or an accurate estimate of σ^2 is available. Without loss of generality, we assume $\sigma^2 = 1$. Write

$$C_p(\hat{\mu}_k) = C_k = \|\mathbf{y} - \hat{\mu}_k\|^2 - n + 2k,$$

where

$$\|\mathbf{y} - \hat{\mu}_k\|^2 = \|\mathbf{y} - \bar{y}\mathbf{1}\|^2 - \sum_{i=1}^{m} \hat{\beta}_i^2 1(|\hat{\beta}_j| > |\hat{\beta}|_{(m-k)}) + k|\hat{\beta}|_{(m-k)}^2$$

We now describe the random walk induced by C_p-Lasso.

For C_p-Lasso,

$$C_{k+1} - C_k = 2 - (k+1)\left(|\hat{\beta}|^2_{(m-k)} - |\hat{\beta}|^2_{(m-k-1)}\right).$$

The final chosen model given by C_p-Lasso will depend on the random walk determined by

$$(k+1)\left(|\hat{\beta}|^2_{(m-k)} - |\hat{\beta}|^2_{(m-k-1)}\right).$$

This random walk is different from those in Woodrofe (1982) and Zhang (1992) due to data induced nested models.

Lemma 3.1 *When $\{V_i^2\}_{1\le i\le m}$ are i.i.d. χ_1^2 random variables,*

(a). $(k+1)\left[V^2_{(m-k)} - V^2_{(m-k-1)}\right] = T_{m-k}/\lambda_{V^2}(a_{m-k})$ where T_i, $1 \le i \le m$, are i.i.d. exponentially distributed random variables with mean 1, $\lambda_{V^2}(x)$ is the hazard function of V^2, and a_{m-k} is between $V^2_{(m-k)}$ and $V^2_{(m-k-1)}$.

(b). For a fixed k, $F_{V^2}(a_{m-k}) \to 1$ almost surely as m goes to ∞.

(c). For a fixed k, $\lambda_{V^2}(a_{m-k})$ converges to 2 almost surely as m goes to ∞.

(d). For $k = pm$, $0 < p < 1$, as $m \to \infty$

$$(k+1)\left(V^2_{(m-k)} - V^2_{(m-k-1)}\right) \xrightarrow{D} \frac{1-p}{f_{V^2}(F^{-1}_{V^2}(1-p))} T_{m-k}$$

and

$$E\left[(k+1)\left(V^2_{(m-k)} - V^2_{(m-k-1)}\right)\right] \to \frac{1-p}{f_{V_1^2}(F^{-1}_{V^2}(1-p))}.$$

(e). ??Suppose that

$$\lim_{x \uparrow x_0} \frac{d}{dx}\left[\frac{1}{\lambda_{V^2}(x)}\right] = 0, \tag{4}$$

where x_0 is ?. For a fixed k,

$$\frac{1 - A_{m-k}}{f(F^{-1}(A_{m-k}))} \to \text{a constant}$$

almost surely as m goes to ∞.

Based on the above Lemma, Under null model or sparse model with strong signals, its performance is similar to the results obtained in Woodrofe (1982) and Zhang (1992) under normal errors. This confirms the comments made by Ishwaran (2004)

> The use of Cp seems to encourage large models in LARS, especially in high-dimensional orthogonal problems,

As in Leng et al. (2006), it cannot be a consistent procedure. However, C_p-Lasso can be improved by increasing its penalty. This is consistent with the suggestion made in Zou et al. (2007) on using BIC instead of AIC. However, increase penalty won't help for abundant models. This suggests that C_p-Lasso is most useful for sparse model.

四、References

Efron, B, Hastie, T. , Johnstone, I., and Tibshirani, R. (2004). Least angle regression (with discussion). *Ann. Statist.* **32**, 407-499.

Fan, J., Huang, T. and Peng, H. (2005). Semilinear high-dimensional model for normalization of microarray data: a theoretical analysis and partial consistency. (with discussion) *J. Amer. Statist. Assoc*., **100**, 781-813.

Leng, C., Lin, Y., and Wahba, G. (2006). A Note on the LASSO and related procedures in model selection. *Statist. Sinica* **16**, 1273-1284.

Opsomer, J.D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, **25**, 186-211.

Mallows, C.L. (1973). Some comments on Cp. *Technometrics* **15**, 661-675.

Stone, C.J. (1985). Additive regression and other nonparametric models. *Ann. Statis*t., **13**, 689-705.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*, Series B **58**, 267-288.

Woodrofe, M. (1982). On Model Selection and the ARC Sin Laws. *Ann. Statist.* **10**, 1182- 1194.

Zhang, P. (1992) On the Distributional Properties of Model Selection Criteria. *J. Amer. Statist. Assoc*. **87**, 732-737.

Zou, H., Hastie, T. and Tibshirani, R. (2007). On the degrees of freedom of the Lasso. *Ann. Statist*. **35**, 2173-2192.

赴國外研究心得報告

計畫編號	NSC 95-2118-M-002-004
計畫名稱	處理微晶片數據之相關多變量分析及模型選取
出國人員姓名 服務機關及職稱	陳宏、國立臺灣大學數學系
出國時間地點	2007/06/23~2007/06/29，University of California at Irvine
國外研究機構	

工作記要：

IMS(數理統計學會)每年舉辦三次研討會，分別在三月、六月及八月。八月的會議規模最大，而三月及六月規模較小，且分別與 International Biometric Society 下之 ENAR 及 WNAR 分別共同舉辦。本人從未參與過 WNAR/IMS 之會議，因 JSM 的議程越來越豐富，與會時反有魚與熊掌不可兼得之慮，故此次特別參與規模較小之 2007WNAR/IMS 之會議。

今年此研討會於美國Los Angeles附近之UC Irvine舉行，含短期課程，會程共有四天，與會者近三百人，規模上與台灣舉辦之會議規模相當。本人在此次會議中，報告了已畢業之碩士生黃信雄之合作論文，“Operating characteristics of Cp-LASSO on variable selection in linear regression with orthonormal regressors”，這被安排在Machine Learning的主題中演講，若與另二位演講相較，就顯得過於理論，她們分別就生物資訊領域之time course data，如何藉由mixture model來進行clustering及在一大型的模擬預測可疑的恐怖份子程式中如何來評估何者會於近期採取攻擊。個人認為如果在台灣舉辦之會議能有此類主題，台灣之統計研究方可再上一層樓。

在本次會議中，另有16篇的邀請博士學生論文，最後亦會推選出最佳論文。在此方面也頗值得台灣統計學界學習借鏡之處。我們雖有最佳論文之選拔，但尚缺一系統性的辦法邀請學生進行報告，不過本年度之南區統計會議，亦已邀請十位國內的博士班學生進行論文報告。

最後本人得由IMS Invited Paper Session: Sparsity in High-Dimensional Problems中理解了此一領域之最近發展，對於個人之未來研究重點及方向助益極大。此次蒙國科會的支持得以成行，能參與此會受益良多，在此謝謝國科會。