行政院國家科學委員會專題研究計畫 成果報告

小樣本多變數模型下貝氏方法之探討
研究成果報告(精簡版)

計 畫 類 別 ：個別型
計 畫 編 號 ：NSC 99-2118-M-009-004-
執 行 期 間 ：99 年 08 月 01 日至 100 年 07 月 31 日
執 行 單 位 ：國立交通大學統計學研究所

計 畫 主 持 人：洪慧念

計畫參與人員：碩士班研究生-兼任助理人員：高君敏
博士班研究生-兼任助理人員：吳侑峻

公 開 資 訊 ：本計畫可公開查詢

中 華 民 國 100 年 10 月 31 日

中文摘要：　近年來統計學家發投入高維度資料的研究，發展出許多方法來分析基因資料。這些資料有著共同的特性就是樣本數不多但是基因數目很多。解決這類的問題，我們分成兩的步驟。首先是如何挑選重要的基因，接著是要如何的利用這些基因做分析。在基因選取方面，我們討論當樣本數固定時，且被測得的基因數目快速增加。倘若影響某疾病的基因數目也固定，我們應該選取多少數目的基因以做資料分析最為恰當。在實際計算時，我們發現計算需要非常耗費時間，並非一時之間可以完成。因此我們從貝氏的角度切入，希望能有一些較簡易的計算方式。在貝氏方法方面，傳統的統計問題因為資料的個數遠大於參數的個數，因此事前分佈的選取變的不是太重要。但在基因的資料方面，往往變數的維度非常的大，
導致參數的維度也非常的大。因此，選取事前分佈變的非常重要。在本計畫中我們考慮當參數的維度很大時，參數空間會以某種形式逼近於一個無母數的空間。這時我們先考慮此無母數空間上的機率測度，然後在進一步考慮此測度所對應於有限個參數上的機率測度。如此應可有效解決問題。

英文摘要：

前言:

近十多年來由於基因晶片的發明產生了大量高密度的 cDNA 陣列資料。因此，不少生物學家，資訊專家及統計學家發投入此類型的研究，發展出許多方法來分析這些基因資料。這些資料有著共同的特性就是樣本數不多但是基因數目很多，亦即所謂的的 large p small n 問題。

研究方法與研究成果:

綜合過去的文獻資料，我們將此問題分成兩的步驟。首先是挑選重要的基因，接著是將這些基因做分析。雖然有非常非常多的研究學者投入此領域的研究，在實務上也發展出許多先進的方法分析此類的資料，有不錯之成果。但在理論方面結果卻不多。在過去的國科會計畫中，我們已經對這前一個問題做了一些有系統性的理論研究。成果如下:在基因選取方面。過去的方法有 (1)在常態分配的假設下，利用 t 統計量選取重要基因 (2) 運用 SVM 的方法選取重要的基因 (3) 最近 Tibshirani 等人在非常分配的假設下提出 SAM 之方法，被廣泛的使用。(4) Tibshirani 等人也提出對於 outliers 做和，然後利用此統計量來做基因選取。另外近幾年也有一些學著嘗試從貝氏的角度來處理這些問題。 George Casella 等人利用適當之事前分配導出一群基因表現量之收縮信賴區間（Shrinkage Confidence Procedures）。W. Jiang 考慮在廣義線性模型下如何利用貝氏方法選取適當之變數。K.E. Lee 等人也將一些簡易之貝氏方法利用在實際的資料上選取重要基因。另外，在資料分析方面，更是百家爭鳴，然而依舊是理論結果並不多。學者提出

關於如何選取恰當的基因數的理論依據，他們選取基因數目的準則是希望分類成功率愈高愈好。幾年前，Bickel 等人證明如果選取太多的基因，不論是用 FisherRule 或是 IndependentRule，在分類上都不會有太好的結果。理論方面，共通的假設是當測量的基因數目愈多時，影響某特殊疾病的基因數目也成一定的方式迅速增多，且觀察的樣本數也以一定的方式增多。在計畫中，我們討論當樣本數固定時，且被測得的基因數目快速增加。倘若影響某疾病的基因數目也固定（或以非常慢的速度增加），我們應該選取多少數目的基因以做資料分析最為恰當。對於這個問題，我們從常態分配的假設下出發，從決策理論的方面著手，探討當一個重要的基因沒有被選取到時的損失為何，同時若我們多選取了一個不會影響疾病的基因來做分類時，會產生誤導而造成的損失又如何。綜合以上兩種損失，（或許再加上每多選一個基因所必須付出的成本）我們能的到一個選取基因數目的恰當方法，能對資料分析有所助益。但在實際計算時，我們發現計算需要非常耗費時間，並非一時之間可以完成。

在這一年內，我們從貝氏的角度切入，希望能有一些較簡易的計算方式。在貝氏方法方面，傳統的統計問題因為資料的個數遠大於參數的個數，因此事前分佈的選取變的不是太重要，只要定義域夠大，結果通常不會太壞。但在基因的資料方面，往往變數的維度非常的大，導致參數的維度也非常的大。因此，選取事前分佈變的非常重要。根據經驗，如果選取不恰當的高維度事前分佈，常常會有很大的偏誤產生。因此，我們考慮當參數的維度很大時，參數空間會以某種形式逼近

於一個無母數的空間。我們先考慮此無母數空間上的機率測度，然後在進一步考慮此測度所對應於有限個參數上的機率測度。在計算上，我們選取的事前分佈會先以計算方便為主要考量。我們採用 Dirichlet Process 事前分佈。由於計畫複雜度頗高，本年度尚無法完成所有工作，未來仍將持續進行。

參 考 文 獻

1. Antoniadis, A., Lambert-Lacroix, S. and Leblanc, F. (2003). Effective dimension reductionmethods for tumor classification using gene expression data. *Bioinformatics* 19 563-570.

2. Buhlmann, P. and Bin, Y. (2004). Discussion of boosting papers. Ann. Statist. 3296–101.

3. Bai, Z. and Saranadasa, H. (1996). Effect of high dimension : by an example of a two sample problem. Statistica Sinica 6, 311-329.

4. Bair, E., Hastie, T., Debashis, P., and Tibshirani, R. (2007) Prediction by supervised principal components. The Annals of Statistics

5. Berry, J. C. (1994). Improving the James-Stein Estimator Using the Stein Variance Estimator. Statist. Prob. Let. 20 241-245.

6. Blackwell, D. (1973). Discreteness of Ferguson selections. Ann. Statist. 1, 356-358.

7. Blackwell, D. and MacQueen, J. B. (1973). Ferguson distributions via Polya urnschemes. Ann. Statist. 353-355

8. Bickel, P. and Levina, E. (2004). Some theory of Fisher's linear discriminant function, naive Bayes, and some alternatives where there are many more variables than observations. Bernoulli 10 989–1010.

9. Boulesteix, A. (2004). PLS Dimension reduction for classification with microarray data. Statistical Applications in Genetics and Molecular Biology 3 1-33.

10. Bura, E. and Pfeiffer, R. M. (2003). Graphical methods for class prediction using dimension reduction techniques on DNA microarray data. Bioinformatics, 19,1252-1258.

11. Cao, H.Y. (2007). Moderate deviations for two sample t-statistics. Prob. and Stat. lett.

12. Chen, S., Donoho, D. and Saunders, M. (2001). Atomic decomposition by basis pursuit. SIAM Rev. 43 129–159.

13. Dettling, M. and Buhlmann, P. (2003). Boosting for tumor classification with gene expression data. Bioinformatics 19 No. 9, 1061-1069.

14. Donoho, D. (2004). For most large underdetermined systems of linear equations of minimal l1-norm solution is also the sparsest solution. 2004-9, Dept. Stat, Stanford Univ.

15. Donoho, D. (2004). For most large undetermined systems of equations, the minimall1-norm near-solution approximates the sparsest near-solution.

16. Dudoit, S., Fridlyard, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American Statistical Association 97 77-87.

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

18. Efron, B. (2006). Minimum Volume Confidence Regions for a Mul-tivariate Normal Mean Vector. J. Roy. Statist. Soc. Ser. B 68 655-670

19. Fan, J and Ren, Y. (2006). Statistical analysis of DNA microarray data. Clinical Cancer Research 12 4469-4473.

20. Fan, J. (1996). Test of significance based on wavelet thresh holding and Neyman's truncation. Journal of the American Statistical Association 91 674-688.

21. Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 1348–1360.

22. Fan, J. and Li, R. (2006). Statistical challenges with high dimensionality: feature selection in knowledge discovery. Proceedings of the International Congress of Mathematicians (M. Sanz-Sole, J. Soria, J.L. Varona, J. Verdera, eds.), Vol. III,595-622.

23. Fan, J. and Lv, J. (2007). Sure independence screening for ultra-high dimensional

24. Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging

number of parameters. Ann. Statist. 32 928–961.

25. Fan, J., Hall, P. and Yao, Q. (2006). To how many simultaneous hypothesis tests can normal, student's t or Bootstrap calibration be applied.

26. Ferguson, T. S. (1973). A Bayesian analysis of some nonparametric problems. Ann Statis. 1, 209-230

27. Ferguson, T. S. (1974). Prior distributions on spaces of probability measures. Ann.Statist. 2, 615-629

28. Ghosh, D. (2002). Singular value decomposition regression modeling for classification of tumors from microarray experiments. Proceedings of the Pacific Symposium on Biocomputing, 11462-11467.

29. Greenshtein, E. (2005). Prediction, model selection and random dimension penal-ties. Sankhy a 67 46–73.

30. Greenshtein, E. (2006). Best subset selection, persistence in high dimensional statistical learning and optimization under l1 constraint. Ann. Statist., 2367-2386

31. Greenshtein, E. and Ritov, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of over parametrization. Bernoulli 10 971-988.

32. Hwang, J. T., Qiu, J. and Zhao, Z. (2008). Empirical Bayes Confidence Intervals Shrinking Both Means and Variances, to appear in J. Roy. Statist. Soc. Ser. B .

33. Hutter, M. (2009). Exact non-parametric Bayesian inference on in_nite trees. Tech.

Rep

34. Zou, H., Hastie, T., and Tibshirani. R. (2007). Outlier sums for differential gene expression analysis, Biostatistics

35. Huang, X. and Pan, W. (2003). Linear regression and two-class classification with gene expression data. Bioinformatics 19 2072-2978.

36. Huber, P. (1973). Robust regression: Asymptotics, conjectures, and Monte Carlo Ann. Statistics 1 799–821.

37. Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric regression. Ann. Statistics. 28 681–712.

38. Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1996). Efficient agnostic learning of neural networks with bounded fan-in. IEEE Trans. Information Theory 42 2118–2132.

39. Lin, Z. and Lu, C. (1996). Limit Theory for Mixing Dependent Random Variables. Kluwer Academic Publishers.

40. Lugosi, G. and Vayatis, N. (2004). On the Bayes risk consistency of regularized boosting methods. Ann. Statistics 32 30–55.

41. Meinshausen, N. (2007). Relaxed Lasso. Computational Statistics and Data Analysis

42. Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable

selection with the Lasso. Ann. Statistics 34 1436–1462.

43. Nemirovski, A. and Yudin, D. (1983). Problem Complexity and Method Efficiency in Optimization. Wiley, New York.

44. Nguyen, D. V. and Rocke, D. M. (2002). Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 18 39-50.

45. Paddock, S. M., Ruggeri, F., Lavine, M., and West, M. (2003). Randomized Polya tree models for nonparametric Bayesian inference. Statist. Sinica 13, 2, 443-460.

46. Pisier, G. (1981). Remarques sur un résultat non publié de B. Maurey. Seminar on Functional Analysis, 1980--1981, École Polytechnic, Palaiseau. Exp. no. V, 13 pp.

47. Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p2/n is large. I. Consistency. Ann. Statistics 12 1298-1309.

48. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. J. Royal Statistics Soc. Ser. B 58 267-288.

49. Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. 99 6567-6572.

50. van der Vaart, A. W. and Wellner, J. A. (1996). Weak Convergence and Empirical Processes. Springer-Verlag, New York.

51. Vapnik, N. V. (1998). Statistical Learning Theory. Wiley, New York.

國科會補助計畫衍生研發成果推廣資料表

日期:2011/10/31

國科會補助計畫	計畫名稱: 小樣本多變數模型下貝氏方法之探討
	計畫主持人: 洪慧念
	計畫編號: 99-2118-M-009-004- 學門領域: 數理統計

無研發成果推廣資料

99 年度專題研究計畫研究成果彙整表

計畫主持人：洪慧念	計畫編號：99-2118-M-009-004-
計畫名稱：小樣本多變數模型下貝氏方法之探討	

成果項目			量化			單位	備註（質化說明：如數個計畫共同成果、成果列為該期刊之封面故事...等）
			實際已達成數（被接受或已發表）	預期總達成數(含實際已達成數)	本計畫實際貢獻百分比		
國內	論文著作	期刊論文	0	0	100%	篇	
		研究報告/技術報告	1	1	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%		
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（本國籍）	碩士生	2	2	100%	人次	
		博士生	1	1	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		
國外	論文著作	期刊論文	0	1	100%	篇	
		研究報告/技術報告	1	1	100%		
		研討會論文	0	0	100%		
		專書	0	0	100%	章/本	
	專利	申請中件數	0	0	100%	件	
		已獲得件數	0	0	100%		
	技術移轉	件數	0	0	100%	件	
		權利金	0	0	100%	千元	
	參與計畫人力（外國籍）	碩士生	2	2	100%	人次	
		博士生	1	1	100%		
		博士後研究員	0	0	100%		
		專任助理	0	0	100%		

其他成果 (無法以量化表達之成果如辦理學術活動、獲得獎項、重要國際合作、研究成果國際影響力及其他協助產業技術發展之具體效益事項等，請以文字敘述填列。)	無

	成果項目	量化	名稱或內容性質簡述
科教處計畫加填項目	測驗工具(含質性與量性)	0	
	課程/模組	0	
	電腦及網路系統或工具	0	
	教材	0	
	舉辦之活動/競賽	0	
	研討會/工作坊	0	
	電子報、網站	0	
	計畫成果推廣之參與（閱聽）人數	0	

國科會補助專題研究計畫成果報告自評表

請就研究內容與原計畫相符程度、達成預期目標情況、研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）、是否適合在學術期刊發表或申請專利、主要發現或其他有關價值等，作一綜合評估。

1. 請就研究內容與原計畫相符程度、達成預期目標情況作一綜合評估

■達成目標

□未達成目標（請說明，以 100 字為限）

□實驗失敗

□因故實驗中斷

□其他原因

說明：

2. 研究成果在學術期刊發表或申請專利等情形：

論文：□已發表 □未發表之文稿 ■撰寫中 □無

專利：□已獲得 □申請中 ■無

技轉：□已技轉 □洽談中 ■無

其他：(以 100 字為限)

3. 請依學術成就、技術創新、社會影響等方面，評估研究成果之學術或應用價值（簡要敘述成果所代表之意義、價值、影響或進一步發展之可能性）(以 500 字為限)

在貝氏方法方面，因變數的維度非常的大，導致參數的維度也非常的大。因此，如果選取不恰當的高維度事前分佈，常常會有很大的偏誤產生。因此，在本計畫中我們考慮當參數的維度很大時，參數空間會以某種形式逼近於一個無母數的空間（維度為無窮大的空間)。這時我們先考慮此無母數空間上的機率測度，然後在進一步考慮此測度所對應於有限個參數上的機率測度。在計算上，我們選取的事前分佈會先以計算方便為主要考量。我們先考慮 Dirichlet Process 事前分佈。此分佈有許多計算上的優點。因此，我們初步的成果有不錯的應用價值，可運用在實際的問題上，未來我們將做更進一步的研究。