行政院國家科學委員會專題研究計畫　期中進度報告

汎函估計之探討(2/3)

計畫類別：個別型計畫
計畫編號：NSC91-2118-M-002-005-
執行期間：91 年 08 月 01 日至 92 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

報告類型：精簡報告
報告附件：國外研究心得報告
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 92 年 6 月 2 日

行政院國家科學委員會補助專題研究計畫期中進度報告

X

汎函估計之探討（2/3）

計畫類別：X 個別型計畫
計畫編號：NSC 91－2118－M－002－005
執行期間：九十一年八月一日至九十二年七月三十一日
執行單位：國立台灣大學數學系

計畫主持人：陳宏

成果報告類型(依經費核定清單規定繳交)：精簡報告

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
　出席國際學術會議心得報告及發表之論文各一份

中　華　民　國　92　年　6　月　2　日

汎函估計之探討（2/3）

一、中英文摘要

不完整數據的統計分析，於近年來極受重視。我們考慮使用 Cronbach α量測測驗中含闕失數據時對信度所造成的影響；另在 cDNA 微陣列所測量的數據，會因染劑、點印基因探針針頭而造成系統性誤差，我們對常用的幾種正規化方法進行比較，並討論這幾種方法間的共同性及差異，並提出一個正規化的新方方法。

關鍵詞：不完整數據、電腦適性測驗、Cronbach α係數、正規化、cDNA 微陣列

Abstract

In recent years, statistical analysis for handling incomplete data has drawn a lot of attentions. In this project, we consider the meaning of Cronbach α when the test includes missing responses. We also consider how to do normalization with cDNA microarray data. It is known that both the efficiency of dye and the gene probes print-tips will introduce systematic bias into the observed measurements. We do a comparison of commonly used normalization methods. In addition, we also propose an alternative normalization procedure.

Keywords: Incomplete data, Computerized adaptive testing, Item response theory, Cronbach α, Normalization, cDNA microarray

二、Incomplete Data, Computerized Adaptive Test, and Normalization in Gene Chip

Missing data has attracted a lot of attention in past twenty years. In this report, we concentrate on two kinds of incomplete data. The first type of data is from computerized adaptive testing (CAT) which has been implemented used in the Graduate Recorded Examination (GRE), the
Graduate Management Admission Test and the National Council Licensure Examination for Nurses in United States for getting a better gauge of examinee's ability. The second one is from normalization of cDNA microarray data. For the first problem, some responses are incomplete (missing) due to design. For the second problem, the data is measured with systematic bias which is being categorized as dye effect, pin effect, and etc.

First, consider CAT in which an individualized test is designed for each examinee. CAT is first proposed in Lord (1970, 1971). The attractive part of CAT is that all examinees may not be required to answer some test items that do not match their abilities. In such a set-up, we are facing the problem of missing data. If we assemble all test items taken by the examinees into a pooled test, each examinee only answer part of the pooled test. The missing pattern is *missing non-ignorable* since it

depends on θ, unknown ability of the examinee under the framework of item response theory.

In our research project, we would like to quantify the *gain* or *loss* of CAT. Suppose the pooled test is administered to all examinees. We would like to quantify the agreement of the test result based on CAT to that based on the pooled test. Specifically, the agreement is measured with Pearson correlation coefficient. However, we never have test score from the pooled test. We apply the method of imputation in which the missing responses of the pooled test are imputed. Then use the imputed score to derive the estimated Pearson correlation coefficient. This kind of agreement measured by Pearson correlation coefficient in education measurement is called Cronbach α. In order to get a deeper understanding of Cronbach α, we finish one technical report on how Cronbach α is affected by the ability of the examinees, the difficulty of test items, and the match of the above two.

Biochips are currently one of the key technologies in Biology. Generally, the arrays that fit these microchips consist of orderly arrangements of samples such as cDNAs, oligonucleotides, or proteins. Macroarraying is the process of organizing sample colonies on large nylon filters to ready them for screening by hybridization. Packing so much information onto a small space gives microarrays a clear advantage. A single DNA microarray plate can contain many thousands of samples, each representing a part of a single gene.

DNA microarray systems are versatile tools for mutational analysis, gene sequencing, and the study of gene expression. The hybridization of nucleic acid–derived samples to the immobilized oligonucleotides in microarrays allows one to easily quantify the expression of specific mRNAs on a genomic scale.

DNA array–based technologies provide relatively simple ways to measure differential gene expression, i.e., the relative levels of RNA transcripts in different cell or tissue samples, for all of the genes of an organism simultaneously. When the levels of specific RNAs from two different sources, such as control and diseased tissue, are measured, their differences can be easily represented if distinctly colored fluorescent labels—blue and yellow, for example—are used to make each sample's cDNA probe. Parallel quantitation of large numbers of mRNA transcripts with the use of microarray technology promises to provide detailed insight into cellular processes involved in the regulation of gene expression (Schena, 1995). DNA microarray information permits researchers to study changes in host-cell gene expression in disease states arising from viral infections or from cell transformation that leads to tumor formation. More complete understanding of these changes should lead to knowledge of mechanisms of virus replication and pathogenesis, as well as host antiviral responses. However, it is well known that systematic variation does exist in the resulting

measurement. The relation between a measured intensity y_{ki} of probe k and the true abundance x_{ki} of molecule type k in sample i may be described as

$$y_{ki} = a_{ki} + b_{ki}\, x_{ki} \quad (1)$$

The gain factor b_{ki} represents the net result of the various experimental effects that come between the count of molecules per cell in the sample and the final readout of the probe intensity, such as: number of cells in the sample, the mean number of label molecules attaching to a sample molecule, hybridization efficiency, label efficiency, and detector gain. The additive term a_{ki} accounts for that part of the measured intensity that does not result from x_{ki}, but from effects such as unspecific hybridization, background fluorescence, stray signal from neighboring probes, and detector offset.

As a first attempt to address this variation, Chen et al. (1997) introduced a decomposition of the multiplicative effect (cf. Eqn. (1)),

$$b_{ki} = b_i\, \beta_k\, (1 + \varepsilon_{ki}). \quad (2)$$

Here, β_k is a probe-specific coefficient, the same for all samples. For each sample i, the normalization factor b_i is applied across all probes. The remaining variation in b_{ki} that cannot be accounted for by β_k and b_i is absorbed by ε_{ki}. Furthermore, since the measured intensities y_{ki} are already "background-corrected" by the image analysis software's local background estimation, Chen et al. assumed the additive effects a_{ki} to be negligibly small. They further simplified the problem in two steps:

First, they noted that one is mainly interested in relative comparisons between the levels of the same gene under different conditions, i. e., in the ratios x_{ki}/ x_{kj}. Hence the probe-specific effects β_k can be absorbed, $\mu_{ki} = \beta_k\, x_{ki}$, simply rescaling the units in which molecule abundances are measured, and need not be determined.

Second, they turned to a stochastic description, and modeled e_{ki} as a normally distributed noise term with mean zero and standard deviation σ, independent of i and k. Thus, in the model of Chen et al. the measured intensity Y_{ki} is a random variable and depends on the true level μ_{ki} as follows:

$$Y_{ki} = b_i\, \mu_{ki}\, (1 + \varepsilon_{ki}), \qquad \varepsilon_{ki} \sim \mathrm{N}(0,\sigma^2) \quad (3)$$

Note that Y_{ki} has constant coefficient of variation σ.

This proposal resolves the issue on this variation. Since then, various methods under the name "Normalization" have been proposed to correct for those systematic bias. All methods rely on the basic assumption that most genes are not differentially expressed. In this project, we do a comparison on major normalization procedure and propose a new procedure.

三、Study on Cronbach α Coefficient

Suppose a test contains I test items. Let Y and Y^* denote the scores when the same test is administered twice and assume memoryless. Denote the score of test item i by X_i and X_i^*. Assume the ability of the examinee is θ. We have

$$\begin{aligned} X_i &= E(X_i|\theta) + \epsilon_i = \tau_i(\theta) + \epsilon_i, \\ Y &= \sum_{i=1}^{I} X_i = \sum_{i=1}^{I} \tau_i(\theta) + \sum_{i=1}^{I} \epsilon_i, \\ Y^* &= \sum_{i=1}^{I} X_i^* = \sum_{i=1}^{I} \tau_i(\theta) + \sum_{i=1}^{I} \epsilon_i^*, \end{aligned}$$

and

$$\begin{aligned} Var(Y) &= \sum_{i=1}^{I} Var(X_i) + \sum_{i \neq j} Cov(E(X_i|\theta), E(X_j|\theta)) \\ &= \sum_{i=1}^{I} Var(X_i) + \sum_{i \neq j} Cov(\tau_i(\theta), \tau_j(\theta)). \end{aligned}$$

When the difficulty levels of all test items are the same, the Pearson correlation coefficient between Y and Y^* is

$$\begin{aligned} \rho_{YY_r} &= \frac{Cov(Y, Y_r)}{\sqrt{Var(Y)}\sqrt{Var(Y_r)}} = \frac{Cov(\sum_{i=1}^{I}(\tau_0(\theta) + \epsilon_i), \sum_{i=1}^{I}(\tau_0(\theta) + \epsilon_i^*))}{Var(Y)} \\ &= \frac{I^2 Var(\tau_0(\theta))}{Var(Y)} = \frac{I^2 \cdot Var(\tau_0(\theta))}{I^2 \cdot Var(\tau_0(\theta)) + I \cdot Var(\epsilon)}. \end{aligned}$$

If we would like to find ρ without knowing Y^*, we need a consistent estimate of $\mathrm{Var}(\tau_0(\theta))$ which can be obtained by $\mathrm{Var}(Y) - \mathrm{Var}(\Sigma_i X_i) = (I^2 - I)\,\mathrm{Var}(\tau_0(\theta))$. We have

$$\begin{aligned} \rho_{YY_r} &= \frac{I(I-1)IVar(\tau_0(\theta))}{(I-1)Var(Y)} \\ &= \frac{I}{I-1}\frac{I^2 Var(\tau_0(\theta)) - I \cdot Var(\tau_0(\theta))}{Var(Y)} \\ &= \frac{I}{I-1}\frac{I^2 Var(\tau_0(\theta)) + \sum_{i=1}^{I} Var(\epsilon_i) - \sum_{i=1}^{I} Var(X_i)}{Var(Y)} \\ &= \frac{I}{I-1}\left(1 - \frac{\sum_{i=1}^{I} Var(X_i)}{Var(Y)}\right). \end{aligned}$$

Therefore, ρ can be obtained based on administering the test once. Moreover, Cronbach α coefficient used in measurement is a measurement on the consistence of repeated testing.

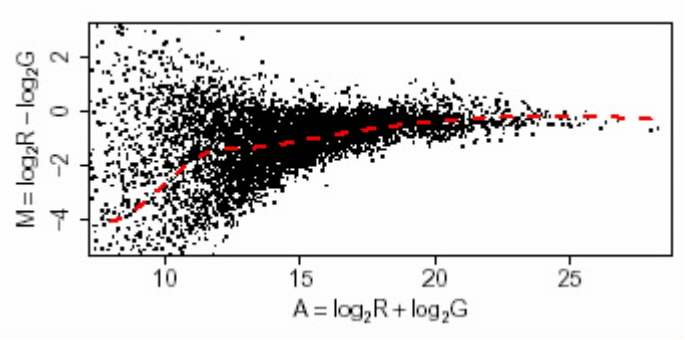
When the difficulty levels of all test items are different, the Pearson correlation coefficient between Y and Y^* is

$$\begin{aligned}
\rho_{YY^*} &= \frac{Cov(Y, Y^*)}{\sqrt{Var(Y)}\sqrt{Var(Y^*)}} \\
&= \frac{Cov(\sum_{i=1}^{I}\tau_i(\theta) + \sum_{i=1}^{I}\epsilon_i, \sum_{i=1}^{I}\tau_i(\theta) + \sum_{i=1}^{I}\epsilon_i^*)}{Var(Y)} \\
&= \frac{\sum_{i=1}^{I} Var(\tau_i(\theta)) + \sum_{i\neq j} Cov(\tau_i(\theta), \tau_j(\theta))}{Var(Y)} \\
&= \frac{\sum_{i=1}^{I} Var(X_i) + \sum_{i\neq j} Cov(\tau_i(\theta), \tau_j(\theta)) - \sum_{i=1}^{I} Var(\epsilon_i)}{Var(Y)} \\
&= \frac{Var(Y) - \sum_{i=1}^{I} Var(\epsilon_i)}{Var(Y)} \\
&= 1 - \frac{\sum_{i=1}^{I} Var(X_i - \tau_i(\theta))}{Var(Y)} = 1 - \frac{\sum_{i=1}^{I} Var(\epsilon_i)}{Var(Y)}.
\end{aligned}$$

Therefore, ρ can no longer be expressed in terms of Y and X_i's only. Moreover, the commonly used Cronbach α coefficient are a bias estimate of Pearson correlation coefficient. A technical report is being prepared to give a detailed discussion on how Cronbach α coefficient is affected by the difficulty level of test items, the ability of the examinees, and whether the test items match with the ability distribution of all examinees.

四、Study on Normalization

Various normalization methods have been proposed in the literature to correct systematic variability in cDNA microarray data analysis. Here we consider the case that two samples are labeled with a green and a red fluorescent dye, respectively. The mixture of the two mRNA preparations is then hybridized simultaneously to a common array on a glass slide. This technology is usually refered to as the Stanford technology (Duggan e al, 1999). We now show a so-called M-A scatterplot of probe intensities in the red and the green color channel from a cDNA array where M = $\log_2$ R - $\log_2$G and A = $\log_2$ R + $\log_2$G.



It is usually assumed that the majority of *genes unchanged*. This scatterplot allows us to assess both measurement noise and systematic biases. Ideally, the data from the majority of the genes that are unchanged should lie on the bisector of the scatterplot. In reality, there are both systematic and random deviations from this.
To adjust for measurement noise and systematic biases, all methods rely on the property that the majority of *genes unchanged*. Some error models have been discussed in Section 1. In addition to it, Kerr et al. (2000) proposed an approach based on the ANOVA technique. They modeled the measured intensity Y_{kjlm} of probe *k* on slide *j*, in the color channel of dye *l*, from a sample that received treatment *m*, as

$$\log Y_{kjlm} = g_k + s_j + d_l + v_m + [gs]_{kj} + [gv]_{km} + \varepsilon_{kjlm}. \qquad (4)$$

As a contrast to Dudoit et al. (2002), they combine the normalization with identification of expressed gene together. Moreover, Dudoit et al. uses loess to find normalization constant which will use less parameters than the ANOVA approach considered in Kerr et al.
A technical report is being prepared to give a detailed discussion on the above two approaches. As a remedy to the reduction of parameters in ANOVA approach, we propose a normalization method based on random effect model.

五、References

[1] Yidong Chen, Edward R. Dougherty, and Michael L. Bittner. (1997) Ratio-based decisions and the quantitave analysis of cDNA microarray images. *Journal of Biomedical Optics*, **2**:364–374.

[2] Sandrine Dudoit, Yee Hwa Yang, Terence P. Speed, and Matthew J. Callow. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**:111–139.

[3] David J. Duggan, Michael Bittner, Yidong Chen, Paul Meltzer, and Jeffrey M. Trent. (1999) Expression profiling using cDNA microarrays. *Nature Genetics*, 21 (Suppl 1):10–14.

[4] M. Kathleen Kerr, Mitchell Martin, and Gary A. Churchill. (2000) Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, **7**:819–837, 2000.

[5] Lord, M. F. (1970). Some test theory for tailored testing. In W.H. Holtzman (Ed.), *Computer-assisted instruction, testing and guidance*. New York: Harper and Row.

[6] Lord, M. F. (1971). Robbins-Monro procedures for tailored testing. *Educational and psychological Measurement*, **31**, 3-31.

[7] Rubin, D. B. (1987). Multiple Imputation for Non-response in Surveys. New York: John Wiley & Sons.

[8] Schena, M. et al. (1995). Quantitative monitoring of gene expression patterns with a cDNA

microarray. *Science* **270**, 467-470.