REPORT DOCUMENTATION PAGE

Form Approved
OMB NO. 0704-0188

1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	May 20, 2002	Final - Jun 1998 thru May 2002

4. TITLE AND SUBTITLE
Statistical Problems in Remote Sensing, Image Compression, and Mapping of Human Chromosomes

5. FUNDING NUMBERS
DAAG55-98-1-0341

6. AUTHOR(S)
Bin Yu

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of California, Department of Statistics
336 Sproul Hall
Berkeley, CA 94720-5940

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U. S. Army Research Office
P.O. Box 12211
Research Triangle Park, NC 27709-2211

10. SPONSORING / MONITORING AGENCY REPORT NUMBER
36606-MA
.18

11. SUPPLEMENTARY NOTES
The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision, unless so designated by other documentation.

12 a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution unlimited.

12 b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Research findings were obtained on Minimum Description Length (MDL) principle and its applications, on micrarray image compression and data analysis, and on classifications based on hyperspectral measurements in remote sensing.

14. SUBJECT TERMS

15. NUMBER OF PAGES
7

16. PRICE CODE

17. SECURITY CLASSIFICATION OR REPORT	18. SECURITY CLASSIFICATION ON THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
UNCLASSIFIED	UNCLASSIFIED	UNCLASSIFIED	UL

Final Report on ARO Grant (DAAG55-98-1-0341)

Principle Investigator: Bin Yu
Statistics Department
367 Evans Hall
University of California
Berkeley, CA 94720-3860

1 Statement of Problems Studied

Under ARO grant (DAAG55-98-1-0341) for the period of June, 1998 through May, 2002, the principle investigator conducted statistical methodology research on Minimum Description Length (MDL) principle and its applications, on microrray image compression and data analysis, and on classification based on hyperspectral measurements in remote sensing.

2 Summary of the Most Important Results

I. Tree recoginition based on hyperspectral measurements

Hyperspectral data consist of intensity readings of hundreds of bands from ground or airborne spectrometers, while multispectral data have typically 4-7 bands. When plotting against the wavelength, a perspectral measurement gives a smooth-looking curve.

I.1 Conifer Tree Species Recognition

Using a high spectral resolution spectrometer, PSD1000 (ANCAL, 1995), measurements were taken at the Blodgett Forest Research Station of the University of California, Berkeley, located on the western slope of the central Sierra Nevada, El Dorado County, California. There are 322 measurements on 6 conifer species with equal proportions: Sugar-pine (SP, Pinus lambertiana), Ponderosa-pine (PP, Pinus ponderosa), White-fir (WF, Abies concolor), Douglas-fir (DF, Pseudotsuga menziesii), Incense-cedar (IC, Calocedrus decurrens), and Giant-sequoia (GS, Sequoiadendron giganteum). After pre-processing, each measurement consists of 179 bands from 350 mn and to 900 nm.

Yu et al (1999) uses hyperspectral measurements collected in the Sierra Nevada Mountains in California to discriminateu six species of conifer trees using a recent non-parametric statis-

tics technique known as Penalized Discriminant Analysis (PDA). A classificati hyperspectral measurements collected in the Sierra Nevada Mountains in California to discriminate six species of conifer trees using a recent non-parametric statistics technique known as Penalized Discriminant Analysis (PDA). A classification accuracy of 76obtained. The emphasis is on providing an intuitive, geometric description of PDA that makes the advantages of penalization clear. PDA is a penalized version of Fisher's Linear Discriminant Analysis (LDA) that can greatly improve upon LDA when there are This discriminative power of hyperspectral data for conifer tree species recognition opens up the possibility of automatic species identification (in contrast to human-involved photo interpretation) based on hyperspectral data for forestry management at a large scale. However, most airborne spectrometers have only several bands operating at a time instead of hundreds as used in the above works. On-going research with M. Hansen, M. Ostland and P. Gong aims at using the above hyperspectral data collected on the ground to select the most discriminative bands to be used with airborne spectrometers. This research uses MDL and other model selection criteria. I.2 Identification of Healthy vs Infected Oak Trees by Sudden Oak Death (SOD)

In Spring 2001, a new project started with the help of a new graduate student Dave Graham-Squire to use hyperspectral measurements for the identification of healthy vs. infected oak trees by Sudden Oak Death, which is a major epidemic threatening oak trees in california and other west coast states.

It is again in collaboration with Prof. Gong's group in ESPM at berkeley. We have so far obtained some preliminary results.

II. Mininum Description Length (MDL) Principle

The PI continued her research program on MDL in the funding period in three different directions.

II.1 A Thorough Review on MDL

Hansen and Yu (2001) focuses on reviewing the field of MDL and pressing for practical applications of MDL such as in regression variable selection. Aiming at exposing MDL to more statisticians, this paper reviews and synthesizes MDL in the context of frequentist and Bayesian statistics, illustrating the connection with real data sets. We make the point that MDL generalizes the maximum likelihood principle of frequentist statistics to model selection problems and it shares many formal derivations of Bayesian statistics. We study three forms of MDL criteria in regression problems, of which one is our close-form new invention. These criteria are investigated in a genetics example, a fruit-classification example using NRC data and a simulation study. It is more than interesting to note that all three MDL criteria automatically select the 'biologically correct' model in the genetics example while AIC or

BIC have to be hand tuned to do so. In general, these MDL criteria are found to behave either like AIC or BIC depending on which is more desirable. Hansen and I are now studying their adaptivity and comparing frequentist and Bayesian procedures in the MDL framework.

II.2 Simultaneous Denoising and Compression via MDL: Adaptive Wavelet Thresholding

With the massive amount of data available in almost every field of statistical applications, the compression aspect of data needs to be addressed formally in statistical inference. Wavelet transformation provides an ideal framework for such a first study. Minimum description length (MDL) criteria are studied in Hansen and Yu (2000) for model selection as flexible forms of thresholding for wavelet denoising and compression. Mixture MDL methods based on a single Laplacian, a two-piece Laplacian, and a generalized Gaussian prior are shown to be adaptive thresholding rules. The MDL procedures achieve mean squared errors comparable with other popular thresholding schemes, but they tend to keep far fewer coefficients. From this property, we demonstrate that our methods represent excellent tools for simultaneous denoising and compression. We make this claim precise by analyzing MDL thresholding in two optimality frameworks; one in which we measure rate and distortion based on quantized coefficients and one in which we do not quantize, but instead record rate simply as the number of non-zero coefficients.

II.3 Other MDL works

Pathwise expansions are obtained in Li and Yu (2000) for the predictive and mixture code lengths used in MDL. The expansions are for exponential families and to the constant order. The results are useful for understanding different MDL forms and provide upper bounds on the Kolmogorov complexity of individual strings.

Rissanen and Yu (2000) is an invited vignette to commemorate Year 2000 by the leading statistics journal JASA. It explains briefly the theoretical pinnings of many information technology products and advoates that theconnection between statistics and information theory is well worth the exploration by statisticians.

III. Microarray image compression and data analysis

A Ph.D. thesis was completed by R. Jornsten under the PI's supervision on microarray image compression and data analysis. It deals directly with the microarray applications, but addresses in general the problem of data compression and its implications for statistical inference. In particular, we consider the following three questions. How can we quantify the effect of compression on statistical inference? How should a compression scheme be designed such that the effect of compression on inference is minimal? How can the Minimum Description Length (MDL) principle be used for model selection with an extraordinary number of

dependent predictors? In this thesis, we attempt to answer these three questions in a general setting, and with a specific application in the compression and analysis of microarray images. The results from the thesis are being published in conference and journal papers as described below.

III.1 Multiterminal Data Compression

In Jornsten and Yu (2002a), we present new results in the context of multiterminal data compression. We derive an improved upper bound on the asymptotic estimation efficiency under rate constraints. Furthermore, we give a geometric interpretation of the new bound, which provides insights into the nature of the multiterminal estimation problem. The bound on asymptotic estimation efficiency gives a gold standard, by which practical compression schemes can be evaluated, and the effect of compression on estimation analyzed.

III.2 Microarray Image Compression

In Jornsten et al (2002), we present a progressive lossy and lossless compression scheme for microarray images. The microarray image technology makes possible the simultaneous measurement of expression levels of thousand of genes. These images have become the standard tools to investigate fundamental biological functions such as gene regulation and interaction, and to discover genetic pathways for diseases such as cancer. They are widely used in laboratories of academia and industry, producing vast quantities of image data. Our compression scheme has been tailored to the microarray image application, such that the essential statistical information in the images is well-preserved at low bit-rates. The compression scheme has a multi-level coded data structure, which allows for fast re-processing and transmission of image subsets.

III.3 Simultaneous Gene Selection and Cluster Analysis via MDL

The information extracted from microarray image experiments provides statisticians with formidable data analysis tasks. In current research, particular attention has been given to the problems of gene clustering and sample classification. Each microarray experiment, or sample, corresponds to a type of tissue, tumor or stage of development. In gene clustering, the goal is to identify genes that exhibit similar expression levels across samples or experiments. In sample classification, a collection of gene expressions is used to build a predictive model for the sample type. In Jornsten and Yu (2002b), we present a new MDL (Minimum Description Length) model selection criterion for the simultaneous clustering of genes, and selection of subsets of gene clusters that function as sample class predictors. For the first time, an MDL selection criterion is given for both predictor variables (genes) and response variables (sample class labels). We are able to build parsimonious classifiers using our MDL model selection criterion that performs better, or as well as the best methods reported in the literature. Our

MDL model selection criterion is generally applicable to prediction problems with highly correlated predictors, where the number of predictors significantly exceeds the number of samples.

3 List of Publications and Technical Reports

Publications:

B. Yu, M. Ostland, P. Gong and R. Pu (1999). Penalized Discriminant Analysis of Hyperspectral Data for Conifer Species Recognition. IEEE Trans. Geoscience and Remote Sensing, vol. 37, p. 2569-2577.

L. Li and B. Yu (2000). Iterated logarithmic expansions of the pathwise code lengths for exponential families. IEEE Trans. Inform. Theory, vol. 46, p. 2683-2689.

J. Rissanen and B. Yu (2000). Coding and compression: a happy union of theory and practice. J. Amer. Statist. Assoc, (Invited Year 2000 Commemorative Vignette), vol. 95, p. 986-988.

M. Hansen and B. Yu (2000). Wavelet thresholding via MDL: simultaneous denoising and compression.", IEEE Trans. Inform. Theory (Special Issue on Information Theoretic Imaging), vol. 46, p.1778-1788.

P. Gong, R. Pu and B. Yu (2001). Conifer species recognition: effects of data transformation. *Int. J. Remote Sensing*, vol. 22(17), 3471-3481.

M. Hansen and B. Yu (2001). Model selection and the principle of Minimum Description Length. J. Amer. Statist. Assoc., vol. 96, p. 746-774.

Conference Papers and Technical Reports:

R. Jornsten and B. Yu (2002a). Multiterminal Estimation: Extensions and a Geometric Interpretation. Proceedings of ISIT2002, Laussanne, Switzerland.

Jornsten and B. Yu (2002b). Simultaneous Gene Clustering and Subset Selection for Sample Classification via MDL. Bioinformatics (special issue on microarray). (submitted).

R. Jornsten, B. Yu, W. Wang, and K. Ramchandran (2002). Compression of cDNA and Inkjet Microarray Images. Proceedings of ICIP2002, Rochester, NY.

4 List of All Participating Scientific Personnel Showing Any Advanced Degrees Earned by Them While Emloyed on the Project

Bin Yu (Principle Investigator)

Rebecka Jornsten (Graduate Student. Ph.D. Completed in Dec. 2001)

Dave Graham-Squire (Graduate Students)