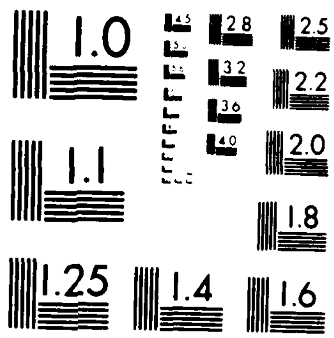
AD-A184 916 SAMPLE REUSE IN STATISTICAL REMODELING(U) SOUTHERN METHODIST UNIV DALLAS TX DEPT OF STATISTICAL SCIENCE W R SCHUCANY AUG 87 N00014-85-K-0340 1/1

UNCLASSIFIED F/G 12/3 NL



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

(12)

A Final Technical Report

For Contract N00014-85-K-0340

Submitted to

OFFICE OF NAVAL RESEARCH

by

SOUTHERN METHODIST UNIVERSITY

Dallas, Texas 75275

Department of Statistical Science

entitled

Sample Reuse in Statistical Remodeling

Prepared by:

William R. Schucany, Ph.D.

Professor of Statistical Science

(214) 692-2464

TABLE OF CONTENTS

I. INTRODUCTION

The nominal period for this contract is the three-year period from 1 June 1985 to 31 May 1988. The Principal Investigators, William R. Schucany and Randall L. Eubank, were supported only during the summer months, June, July and August, for the calendar years 1985, 1986, and 1987. Thus the funded efforts are essentially complete at the date of this report. If we delayed this report until the end of the contract period, it is likely that some additional technical reports would be published or accepted. Rather than postpone the Final Report, it is being submitted now so that there might be a timely consideration of the accompanying renewal proposal.

The primary objective of this contract has been the development and extension of sample reuse techniques in statistical modeling. To a great extent the models that have been examined are nonparametric or non-Gaussian. The resampling schemes that have been employed and extended are the jackknife, the bootstrap and generalized cross validation. Many new and potentially important results have been established. These are summarized in the next section. In Section IV our technical proposal presents a continuation of the discussion of some of these topics. In that section we outline the objectives of our research and the approaches to be taken during the proposed two-year renewal of the project.

II. RESEARCH ACCOMPLISHMENTS

A. Splines and Nonparametric Regression

A number of new results concerning diagnostic and inferential analysis have been developed during the project period. A general treatment of diagnostic analysis for penalized least-squares estimators was provided in Eubank and Gunst (1986) through the extension of work by Eubank (1985a). In the Eubank (1985a) paper it was shown that diagnostics patterned after those from linear regression analysis were suitable for use with smoothing splines as well. The Eubank and Gunst paper established that this was true for penalized least-squares estimators in general, not just smoothing splines alone. This has the consequences of providing a plethora of diagnostic tools that can be used with such methods as ridge regression, nonlinear regression via Marquardt's algorithm, multivariate and partial smoothing splines, method of regularization estimators, etc. Carmody (1985) gave a detailed treatment of the Eubank/Gunst proposals for multivariate smoothing splines in his Ph.D. dissertation.

A new method of constructing robust penalized regression estimators has been studied by Cunningham (1987) in his Ph.D. work. In contrast to other proposed methods, Cunningham's approach provides for simultaneous estimation of both location and scale parameters. The estimation method also has diagnostic implications since the amount of downweighting given an observation provides an indication as to whether or not it should be viewed as an outlier.

Eubank gave an overview of the diagnostic developments mentioned above at Ohio State University in March of 1986. This was an invited presentation for the NSF conference on "Spline and Partial Spline Methods in Statistics" featuring Grace Wahba as the principal speaker.

In Eubank (1986a) a connection was drawn between several types of Bayesian nonparametric regression and spline smoothing in partially linear models. A posterior covariance kernel was derived for partial smoothing splines that makes it possible to extend Bayesian prediction intervals, which are currently used with ordinary smoothing splines, to the partial spline setting.

A treatment of the general subject of nonparametric regression is given in Eubank (1987b). This text contains an account of both the theory and practice of nonparametric regression slanted towards spline smoothing issues.

Accomplishments in Data Modeling

Several new developments during the contract period were concerned with quantile based methods for statistical data modeling. Eubank and LaRiccia (1986, 1987) developed analogs of the Wald test for parametric hypotheses based on sums of squared L-statistics. These tests are based on either the entire set, or a subset, of the order statistics and will have the same power as the likelihood ratio test under local alternatives. However, like the Wald test, no parameter estimation is required. Two review articles (Eubank (1985b, 1986b)) on quantiles and the related problem of optimal spacing selection were published in the Encyclopedia of Statistical Sciences.

A new approach to hypothesis testing was developed by Eubank, LaRiccia and Rosenstein (1987). The method is based on the fact that hypothesis testing problems are really distribution comparison problems. This allows for the construction of a comparison density that must be uniform under the hypothesis of interest. The squared error distance between the comparison density and the uniform density can be broken into components which can be easily estimated from data and must vanish under the hypothesis. Many classical tests can be derived from this perspective including, e.g., the t-test, the Wilcoxon signed rank and rank sum tests, the Kruskal-Wallis test, the F-test for ANOVA, Pearson's product moment correlation coefficient, Spearman's ρ, etc. However, there are many more new tests whose properties have as yet to be explored. Rosenstein (1986) studied the properties of some of these new tests in the context of one-sample tests for symmetry.

B. Bootstrap Methodology

The two major objectives for the research effort on this topic were refinements of confidence intervals and applications in time series. There have been significant accomplishments on both.

An important tool in the understanding and further development of many statistical procedures is the influence curve related to a statistical functional, $T(F_n)$. The results in Michael and Schucany (1985) extend the previous work in this area to goodness-of-fit statistics. For the large class of problems in which the test is based upon a functional at the empirical distribution function, this work made a new connection between influence curves and classical measures of efficiency due to Pitman and Bahadur. Many other results in goodness-of-fit for censored samples were compiled in the chapter by Michael and Schucany (1986).

With this foundation of knowledge about influence functions, refinements to bootstrap confidence intervals suggested by Efron (1987) and commented upon by Schucany (1987a) could be further refined and extended. Another general background article on resampling results by Schucany (1987b) ties together jackknife, bootstrap, cross validation and rerandomization. The key ingredient in most current approaches to improving the small sample behavior of techniques such as the jackknife and bootstrap, is an expansion of the functional, $T(F_n)$, or of its distribution function or both. Frangos and Schucany (1987a) used the jackknife, as the finite-sample first-order influence function $T(F_n)$, to obtain a one-term expansion that performed as well as Efron's (1987) accelerated bootstrap. In the same report Frangos and Schucany demonstrated the small sample superiority of that approach over the proposals that take higher order terms of an Edgeworth expansion into account.

In a second report Frangos and Schucany (1987b) examined the small sample performance of intervals that utilize second-order influence functions to yield a refined approximation to the statistical functional. In extensive simulation experiments involving variance and correlation functionals, several distributions and sample sizes, the second-order method provided further improvement of the actual coverage over that obtained by the accelerated bootstrap. An additional important finding in this work is that the bootstrap of studentized quantities yields consistently better results than any of the accelerated versions.

In another investigation of variance estimators for U-statistics Schucany and Bankson (1987) compared the ordinary jackknife to newly proposed estimators. In small samples they demonstrated the non-negligible character of second-order terms. After introducing new unbiased estimators for those terms, they showed that the jackknife successfully captures these contributions without the additional computational work required for direct estimation.

Some potentially important results were obtained by applying the bootstrap to the problem of constructing prediction intervals for stationary autoregressive processes in non-Gaussian cases. Thombs and Schucany (1987) proposed a technique for resampling the residuals process and generating bootstrap replications from the relevant conditional distribution of realizations that have the same fixed values at the end of the series. This application of the bootstrap principle produced a nonparametric estimate of the quantiles of the conditional distribution of future values of the autoregressive process. Simulations established the potential effectiveness of this approach to forecasting.

References other than those appropriately listed in Section III

Carmody, Thomas J. (1985), Diagnostics for Multivariate Smoothing Splines. Doctoral Dissertation, Dept. of Statist., Southern Methodist Univ.

Cunningham, Kelly (1987), Robust Penalized Regression. Doctoral Dissertation, Dept. of Statist., Southern Methodist Univ.

Efron, B. (1987), "Better bootstrap confidence intervals," J. Amer. Statist. Assoc., 82, 171-185.

Rosenstein, R.B. (1986), Components of Phi-Squared and Tests of Symmetry. Doctoral Dissertation, Dept. of Statist., Southern Methodist Univ.

Thombs, L.A. (1985), Bootstrap Prediction Intervals for Autoregressive Processes. Doctoral Dissertation, Dept. of Statist., Southern Methodist Univ.

III. PAPERS PUBLISHED AND PRESENTED

The listing of reports in this section are categorized in the format specified for the two previous annual summaries that were submitted for this contract.

a. Papers Submitted to Refereed Journals (not yet published)

Eubank, Randall L. (1987a), "Optimal Grouping, Spacing, Stratification and Piecewise Constant Approximation", to appear SIAM Review.

Eubank, Randall L. and LaRiccia, Vince (1987), "Regression Type Tests for Parametric Hypotheses Based on Optimally Selected Subsets of Order Statistics," submitted to Annals of Statist. Math.

Eubank, Randall L., LaRiccia, Vince and Rosenstein, R.B. (1987), "Some New and Classical Test Statistics Derived as Components of Pearson's Phi-squared Distance Measure," to appear J. Amer. Statist. Assoc.

Thombs, Lori A. and Schucany, William R. (1987), "Bootstrap Prediction Intervals for Autoregression," submitted to J. Amer. Statist. Assoc.

b. Papers Published in Refereed Journals

Eubank, Randall L. (1985a), "Diagnostics for Smoothing Splines," J. Roy. Statist. Soc., B, 47, 332-341.

Eubank, Randall L. (1986a), "A Note on Smoothness Priors and Nonlinear Regression". J. Amer. Statist. Assoc., 81, 514-517.

Eubank, Randall L. and Gunst, Richard F. (1986), "Diagnostics for Penalized 'east-Squares Estimators", Statist. and Prob. Letters, 4, 265-272.

Eubank, Randall L. and LaRiccia, Vince (1986), "Regression Type Tests for Parametric Hypotheses Based on Sums of Squared L-Statistics" J. Statist. Planning and Infer., 14, 401-407.

Eubank, Randall L. and Webster, J.T. (1985), "The Singular-Value Decomposition as a Tool for Solving Estimability Problems in Statistics," Amer. Statist., 39, 64-66.

Michael, John R. and Schucany, William R. (1985), "The Influence Curve and Goodness of Fit," J. Amer. Statist. Assoc., 80, 678-682.

Schucany, William R. (1987a), "Comment on 'Better Bootstrap Confidence Intervals' by B. Efron", J. Amer. Statist. Assoc., 82, 196-197.

c. <u>Books (and sections thereof) submitted for publication</u>:

Eubank, Randall L. (1987b), <u>Spline Smoothing and Nonparametric Regression</u>, Marcel Dekker, Inc.

Schucany, William R. (1987b), "Sample Re-Use" in <u>Encyclopedia of Statistical Sciences</u>, Wiley

d. <u>Books (and sections thereof) published</u>:

Eubank, Randall L. (1985b), "Optimal Spacing Problems," in the <u>Encyclopedia of Statistical Sciences</u>, <u>6</u>, 452-458, Wiley.

Eubank, Randall L. (1986b), "Quantiles", in <u>Encyclopedia of Statistical Sciences</u>, <u>7</u>, 424-432, Wiley.

Michael, John R. and Schucany, William R. (1986), "Analysis of Data from Censored Samples", Chapter 11 in <u>Goodness-of-Fit Techniques</u>, ed. by R. D'Agostino and M. Stephens, Marcel Dekker, Inc.

g. <u>Invited Presentations at Technical Society Conferences</u>

William R. Schucany, "Using the Jackknife and Bootstrap Cautiously" (one hour) invited by Cincinnati Chapter of ASA, Conference on Statistics, October 5, 1984.

William R. Schucany, "Minimum Distance Estimation: Tradeoff Between Efficiency and Robustness" (one hour) invited by Cincinnati Chapter of ASA, Conference on Statistics, October 5, 1984.

William R. Schucany, "Statistical Issues in Employment Discrimination Litigation" (one hour) invited by Florida Chapter of ASA, Orlando, Florida, March 1, 1985.

Randall L. Eubank, "Parameter Estimation from Randomly Censored Data" (35 minutes) invited talk at the IMS regional meeting, Austin, Texas, March, 1985.

Randall L. Eubank, "Discussion of Papers by Speckman and Marron on Nonparametric Regression" Summer Research Conference/Southern Region Comm. on Statistics, Mobile, Ala., June 20, 1986.

Randall L. Eubank, "Diagnostics for Penalized Least Squares Estimators", (one hour) invited talk at NSF/CBMS Conference Ohio State University, March, 1986.

h. <u>Contributed Presentations</u>

Randall L. Eubank, "Comparison Densities and Components of Phi-squared" Joint Statistical Meetings, Chicago, Ill., August, 1986.

Lori A. Thombs and William R. Schucany, "Prediction in Autoregression Using the Bootstrap", Joint Statistical Meetings, Chicago, Ill., August, 1986.

j. Technical Reports

Schucany, William R., Thombs, Lori A., and Cunningham, Kelly (1986), "Generating Jointly Distributed Variates by Restricted Random Sampling", Tech. Rpt. No. SMU-DS-TR-197, (supported by Sandia National Laboratories)

Frangos, C.C. and Schucany, William R. (1987a), "Jackknife-inspired Improvements of Bootstrap Confidence Intervals," Tech. Rpt. No. SMU-DS-TR-204.

Frangos, C.C. and Schucany, William R. (1987b), "Bootstrap Confidence Intervals Using Influence Functions," Tech. Rpt. No. SMU-DS-TR-205.

Schucany, William R. and Bankson, Daniel M. (1987), "Small Sample Variance Estimators for U-Statistics," Tech. Rpt. No. SMU-DS-TR-206.

IV. OTHER ACTIVITIES

In addition to directing the four doctoral dissertations, which are listed at the end of Section II, the principal investigators were engaged in various other professional activities that would not otherwise be listed among the research accomplishments.

William R. Schucany

Associate Editor, J. Amer. Statist. Assoc., until 1986
Associate Editor, Commun. in Statist., to present
Referee for: JASA, Commun. in Statist., Technometrics
Chair, ASA Section of Statistical Education, 1986
ASA Committee on Economic Status of the Profession, present
Invited presentation, "Resampling Methodology", to ONR/NRL-Tech-410 Conference, Sept. 26, 1986

Randall L. Eubank

Associate Editor, Amer. Statistician, 1987
Associate Program Secretary, IMS Central Region, 1986 to present
Referee for: JASA, Annals of Statist., Commun. in Statist.

END
10-81
DTIC