Paper 254-27

SAS© Implementations of Nonparametric Smoothing and Lack-of-Fit Tests Based on Smoothers

George F. von Borries, Texas A&M University, College Station, TX
Jeffrey D. Hart, Texas A&M University, College Station, TX

1 Abstract

A user-friendly program in SAS® is proposed to perform many nonparametric smoothing and testing procedures. The nonparametric part of the program consists of resources already available on SAS and the testing part will consist of lack-of-fit tests based on smoothing ideas that will be continuously updated. Four main procedures are described for the testing part of the program: tests of a linear model based on a ratio of quadratic forms and on probability bands, the "order selection" test of fit of a linear model, and a test of fit of a nonlinear regression model. The user-friendly characteristic of this program requires basic knowledge of the *SAS System* using the SAS Macro language as the base for implementation with modules SAS/STAT®, SAS/IML®, and SAS/GRAPH®.

2 Introduction

Version 8.2 of the *SAS System* presents many procedures for nonparametric density estimation and nonparametric regression. SAS/STAT includes the procedures `KDE` for kernel density estimation, `LOESS` for local regression, `TPSPLINE` for estimation of multivariate regression surfaces with thin-plate smoothing splines, and `GAM`, which contains a powerful array of smoothing tools for fitting generalized additive models. In addition to the procedures in SAS/STAT, SAS/INSIGHT® also provides nonparametric curve-fitting using smoothing splines, kernel density estimates, loess and fixed bandwidth local polynomial estimators. Recently SAS/IML® implemented a group of macros for wavelet analysis. Smoothers based on wavelets are used to nonparametrically estimate functions. In addition to all these resources, the *SAS System* has the subroutine FFT that computes a fast Fourier transform, thus allowing one to obtain Fourier series estimators.

The smoothing techniques available in SAS constitute powerful tools for data analysis and research in the area of nonparametric regression. However, there remains some very useful methodology that has not yet been incorporated into either SAS/STAT or SAS/INSIGHT, one example being smoothing-based lack-of-fit tests as described in Hart (1997). Among other things, these methods can be used to test the fit of parametric regressions, the additivity of a regression model and the assumption of homoscedasticity. We are not aware of any widely available statistical package that has implemented such procedures.

The authors are developing a user-friendly program using the SAS Macro Language that will perform many of the nonparametric smoothing and testing procedures presented in Hart (1997). The key resources for the user-friendly part of this program are the use of the `%Window` macro command to comunicate with the user and the *Output Delivery System* to produce outputs in `pdf`, `html` and `rtf` formats. At this moment the nonparametric smoothing part

of the program consists of some macros that create an interface for use of available resources on *SAS Software*. The testing part of the program, which is to be continually updated, will consist of lack-of-fit tests based on smoothing ideas. This paper describes four main procedures developed for the testing part of the program:

- Test of fit of a linear model based on a ratio of quadratic forms
- "Order selection" test of fit of a linear model
- Test of fit of a linear regression model based on probability bands
- Test of fit of a nonlinear regression model

The rest of the article describes the basics of these procedures; further details may be found in Hart (1997). The program will use several modules of the *SAS System*, including SAS/STAT, SAS/IML®, SAS/GRAPH and the SAS Macro language. Resources available on SAS/INSIGHT should be also included. In addition to the four procedures above, we have plans for implementing likelihood-based and Bayesian nonparametric tests of model fit. Updated information and new versions of the program will be available at the web sites `http://stat.tamu.edu/~hart` and `http://stat.tamu.edu/~gborries`.

3 Test of Adequacy of a Linear Regression Function

We describe three procedures for testing the fit of a linear regression model. To facilitate our discussion we assume that the hypothesis to be tested is

$$H_0 : r(x) = \theta_0 + \theta_1 x, \quad \text{for all } x, \tag{1}$$

where r is the underlying regression function. We assume throughout that the observed data are independent and of the form $(x_1, Y_1), \ldots, (x_n, Y_n)$, where the covariate values $x_1, \ldots, x_n$ are either fixed or conditioned upon. For convenience we assume the x_i's lie in the interval $[0, 1]$.

3.1 Tests based on ratios of quadratic forms

Here we describe an alternative approach to testing a null hypothesis of the form

$$H_0 : r(x) = \sum_{j=1}^{p} \theta_j r_j(x). \tag{2}$$

Let $\boldsymbol{e}$ be a column vector of ordinary least squares residuals obtained from fitting the linear function in (2). Consider a linear smoother with weights $w_j(x)$, $j = 1, \ldots, n$. If W is the smoother matrix with ijth element $w_j(x_i)$, then a test statistic for hypothesis (2) is

$$R = \frac{\boldsymbol{e}'W'W\boldsymbol{e}}{\boldsymbol{e}'C\boldsymbol{e}}, \tag{3}$$

where C is a matrix not depending on the data and having the property that $\boldsymbol{e}'C\boldsymbol{e}$ is an unbiased (or nearly unbiased) estimator of the variance of Y_i.

The statistic R is a smoother-based analog of an F-statistic for testing (2). The null hypothesis is rejected when R is "large." Two procedures are implemented for approximating a P-value. One is described in Section 5.4.2 of Hart (1997) and is based on the assumption that the data are normally distributed. The second procedure is a bootstrap algorithm described in Section 5.4.3 of Hart (1997).

3.2 An order selection test

Another test of (2) is discussed in Chapter 7 and Section 8.2 of Hart (1997). This test may be described as follows:

1. Apply a "truncated" Fourier series smoother to residuals from the fitted linear model. Truncated refers to the fact that the series is truncated after a certain number of basis functions have been included in the series.

2. Use a modified Mallows-like criterion to choose the truncation point of the series smoother.

3. Reject the null hypothesis if one or more terms of the series are chosen in step 2.

The modification of the Mallows criterion referred to in step 2. ensures that the size of the test is approximately equal to a nominal level. Our implementation of the order selection test approximates a P-value in two ways: on the assumption of normality and by use of a bootstrap procedure. We will also display the result of the test graphically. The order selection test has the appealing property that the null hypothesis is "accepted" if and only if the series smoother applied to the residuals is perfectly flat. Hence, by showing a plot of the series smoother, the outcome of the test is evident.

It should be noted that, while the description of this section has been confined to simple regression, our software will allow the user to test the fit of a linear model in *multiple regression* as well.

3.3 Approach based on probability bands

Probability bands can be used to test hypothesis (1), where θ_0 and θ_1 are unknown parameters. Using a smoother $\hat{r}(x)$ to estimate the regression function, and the least squares estimates $\hat{\theta}_0$ and $\hat{\theta}_1$ of the parameters θ_0 and θ_1, define

$$\delta(x) = \hat{r}(x) - \hat{\theta}_0 - \hat{\theta}_1 x. \tag{4}$$

The bootstrap test consists of fitting the straight line, obtaining the residuals $e_i = Y_i - \hat{\theta}_0 - \hat{\theta}_1 x_i$, $i = 1, \ldots, n$, and then using the empirical distribution of the e_i's to construct "null" probability bands. These bands approximate ones having the following property: when H_0 is true, a properly standardized version of $\delta(x)$ lies completely within the bands with probability $1-\alpha$. The null hypothesis is thus rejected if the observed, standardized δ ever strays outside the bootstrap bands.

A P-value is computed by using the bootstrap simulation to estimate

$$P_{H_0}\left(\max_{1\le i\le n} \frac{|\delta(x_i)|}{\hat{\sigma}(\delta)} \ge u\right), \tag{5}$$

in which P_{H_0} denotes probability as computed under H_0 and $\hat{\sigma}(\delta)$ the standard deviation of $\delta(x)$.

Hart (1997) uses a Gasser-Müller estimator as the smoother $\hat{r}(x)$. Output from the program consists of a bootstrap P-value and a graph displaying the null probability bands and the observed $\delta(x)/\hat{\sigma}(\delta)$, $0 \le x \le 1$.

4 Testing The Fit of a Nonlinear Model

An iterated bootstrap procedure is used to test for lack of fit of a nonlinear model

$$Y_i = r_\theta(x_i) + \sigma\epsilon_i, \quad i = 1, \ldots, n, \tag{6}$$

where $\epsilon_1, \ldots, \epsilon_n$ are i.i.d with mean 0, variance 1 and common distribution function F. The iterated bootstrap procedure is applied because lack-of-fit test statistics, (3) for example, are typically not pivotal quantities when the null model is nonlinear. When a test is not based on a pivotal quantity, approximating its level by that of a large-sample or single bootstrap test can be quite inaccurate, as discussed in Hall (1992).

Let $P(A|(\theta_0, \sigma_0, F_0))$ denote the probability of some event A determined by data from the regression model (6), and θ_0, σ_0 and F_0 be the true values of θ, σ and the distribution of ϵ_i, respectively. If S_n is a statistic of the form (5), our objective is to approximate c, where

$$P(S_n \ge c|(\theta_0, \sigma_0, F_0)) = \alpha. \tag{7}$$

In general c depends on θ_0, σ_0, and F_0. Letting $\hat{F}$ denote the empirical distribution of the standardized residuals $\hat{\epsilon}_1, \ldots, \hat{\epsilon}_n$, the iterated bootstrap estimate

involves obtaining the value of t, call it t_0, that solves the equation

$$P(S_n \geq tc(\hat{\theta}, \hat{\sigma}, \hat{F})|(\theta_0, \sigma_0, F_0)) = \alpha.$$

The quantity t_0 is estimated by the solution to the equation

$$P(S_n^* \geq tc(\hat{\theta}^*, \hat{\sigma}^*, \hat{F}^*)|(\hat{\theta}_0, \hat{\sigma}_0, \hat{F}_0)) = \alpha, \quad (8)$$

where S_n^*, $\hat{\theta}^*$, $\hat{\sigma}^*$ and $\hat{F}^*$ are functions of data randomly generated from model (6) with $\theta = \hat{\theta}$, $\sigma = \hat{\sigma}$ and $F \equiv \hat{F}$. To approximate $P(S_n \geq tc(\hat{\theta}, \hat{\sigma}, \hat{F})|(\theta_0, \sigma_0, F_0))$ for a given t, B_1 bootstrap samples are generated from the model $(\hat{\theta}, \hat{\sigma}, \hat{F})$ and in the ith bootstrap sample, $i = 1, \ldots, B_1$, B_2 bootstrap samples are generated from the model $(\hat{\theta}_i^*, \hat{\sigma}_i^*, \hat{F}_i^*)$, leading to a total of $B_1 B_2$ samples of size n. Approximation of $c(\hat{\theta}_i^*, \hat{\sigma}_i^*, \hat{F}_i^*)$ is done by means of the B_2 samples at the second stage. Having obtained $c(\hat{\theta}, \hat{\sigma}, \hat{F})$ and $\hat{t}_0$, the solution of (8), the null hypothesis is rejected at level α if S_n exceeds $\hat{t}_0 c(\hat{\theta}, \hat{\sigma}, \hat{F})$.

The above iterated bootstrap procedure parallels that described in Hall (1992, pp. 20-22). It will be implemented using smoothers available in SAS/STAT and output from `PROC NLIN`.

5 Plans for Expanding The Software

A recent paper of Aerts, Claeskens, and Hart (1999) (ACH) generalizes the order selection test described in Section 3.2. In many settings one has a likelihood depending on an unknown function of interest, call it g, and finitely many nuisance parameters. For such settings, ACH propose a very general version of the order selection test for checking the fit of a parametric model for g. An important setting in which this test can be applied is checking the fit of a generalized linear model. Eventually, we intend to implement the ACH test as part of our software.

Another method we plan to implement is the Bayes order selection test described in Section 7.6.5 of Hart (1997). This procedure uses noninformative priors for all unknown aspects of the model, and computes a posterior probability of the hypothesis that a response and a covariate are unrelated.

6 Conclusion

The *SAS System* has excellent tools to produce high quality user-friendly programs, as for example the resources available in the SAS/AF® module. With the simple tools available in the SAS Macro Language, it is also possible to produce good user-friendly programs. Our program intends to put together the many tools available in the *SAS System* for nonparametric density estimation and regression and also to include new routines for lack-of-fit tests based on smoothing ideas. The program will be continually updated and revised and is not intended as a substitute for the procedures of the *SAS System*, which have the advantage of "programming flexibility." Our program will provide easy-to-use tools for doing smoothing research and data analysis.

References

[1] Aerts, M., Claeskens, G., Hart, J.D. "Testing the Fit of a Parametric Function," *Journal of the American Statistical Association*, 94, 869-879, 1999.

[2] Hall, P. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, New York, 1992.

[3] Hart, J.D. *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer-Verlag, 1997.

Contact Information

George F. von Borries is a graduate student, Department of Statistics, Texas A&M University, sponsored by CNPq, Brazil. He is also assistant professor (on leave), Universidade de Brasília, Brazil.

Jeffrey D. Hart is Professor, Department of Statistics, Texas A&M University. He is a Fellow of both the American Statistical Association and the Institute of Mathematical Statistics. He has won an outstanding teaching award at Texas A&M University, authored or co-authored over 40 published papers, and is author of the book *Nonparametric Smoothing and Lack-of-Fit Tests*.

Your comments and questions are valued encouraged. Contact the authors at:

George F. von Borries, Ph.D. Student

Office:	506 Blocker Building College Station - TX 77843 USA
Phone:	(979) 731-1306
Fax:	(979) 845-3144
E-Mail:	gborries@stat.tamu.edu
Web:	stat.tamu.edu/~gborries

Jeffrey D. Hart, Professor

Office:	404D Blocker Building College Station - TX 77843 USA
Phone:	(979) 845-3178
Fax:	(979) 845-3144
E-Mail:	hart@stat.tamu.edu
Web:	stat.tamu.edu/~hart