

SASMLF - 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

SASMLF is a user-friendly program written with SAS® that 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. Two main procedures are described for the testing part of the program: tests of a linear model based on ratios of quadratic forms and the 'order selection' test of fit of a linear 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, *TP-SPLINE* 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 also 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 [6]. 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 in SAS Macro Language that will perform many of the nonparametric smoothing and

testing procedures presented in Hart [6]. The key resources for the user-friendly part of this program are the use of the *%Window* macro command to communicate 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 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 two main procedures developed for the testing part of the program:

1. Test of fit of a linear model based on a ratio of quadratic forms
2. "Order selection" test of fit of a linear model

Section 3 describes the basics of these procedures; further details may be found in Hart [6]. Section 4 describes the basic tools used to develop the "user-friendly" structure of the program *SASMLF*.

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 also be included. In addition to the two tests above, we have plans for implementing test of fit of a linear regression model based on probability bands, test of fit of a nonlinear regression model and Bayesian nonparametric tests of model fit. Updated information and new versions of the program will be available at the web site <http://stat.famu.edu/~gborries/sastips.html>.

3 Test of Adequacy of a Linear Regression Function

We describe two procedures for testing the fit of a linear regression model. We assume throughout that the observed data are independent and of the form $(x_1, Y_1), \dots, (x_n, Y_n)$, where the covariate values $x_1, \dots, x_n$ 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

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

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

$$R = \frac{e'W'We}{e'Ce}, \quad (2)$$

where C is a matrix not depending on the data and having the property that $e'Ce$ 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 [6] 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 [6].

3.2 An order selection test

Another test of (2) is discussed in Chapter 7 and Section 8.2 of Hart [6]. 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 as described in Hall [5]. 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.

4 Basic Tools for a User-Friendly Program and SASMLF Macro

The SAS System has development tools such as *SAS/AF*, *Screen Control Language* and *SAS/EIS*[®] to produce high quality user-friendly programs. SAS Guide to Applications Development [3] introduces the user to all these tools.

A simple way to implement "friendly" programs in SAS is to use SAS Macro Language. It is possible to produce good user-friendly programs using the *%Window* macro command and *Macro Autocall Facility* combined with *Output Delivery System* and *Pmenu Procedure*. Our experience is that all these tools are usually well accepted by both technical and non-technical staff. The advantage is that they require basic knowledge of SAS programming in contrast with the more sophisticated ones like *SAS AF*.

The main characteristic of a user friendly program is to interact with the user in a way they can receive and return information to the application. The *%Window* macro command provides a simple way to do that. As an example Figure 1 has part of the main window of *SASMLF*. This window has some basic elements of the *%Window* command such as underlined and highlighted text and places to input commands.

The following code is used to produce the window in Figure 1.

```
data SASMLP;
  %WINDOW WELCOME IROW=1 ROWS=50 color=white
  #3 @25 'TEXAS A&M UNIVERSITY - DEPARTMENT OF STATISTICS'
    attr=underline color=red
  #5 @26 "NONPARAMETRIC SMOOTHING AND LACK-OF-FIT TESTS"
    attr=underline color=red
  #10 @15 "-----"
    "-----" color=red
  #12 @15 "Study Code:" color=red
    @27 stdcod 13 attr=underline persist=no color=red
    @42 "Simulation #:" color=red
    @57 simnum 3 attr=underline persist=no color=red
  #14 @15 "-----"
    "-----" color=red
  #16 @15 "SAS Programs for Nonparametric Smoothing:"
    color=red attr=rev_video

  #18 @15 MAC1 1 attr=underline persist=no color=red
    @18 "1. Local Linear Regression (LOESS)" color=red
  #20 @18 "2. Wavelet Analysis (IML)"
    attr=rev_video color=blue
  #22 @15 "-----"
    "-----" color=red
  #24 @15 "Lack-of-Fit Tests Based on Smoothers"
    color=red attr=rev_video
  #26 @15 TES1 1 attr=underline persist=no color=red
    @18 "1. Order Selection Test (Linear Model)" color=red
  #28 @18 "2. Nonlinear Regression Model" color=blue
    attr=rev_video
  #31 @15 MACEND 1 attr=underline persist=no color=red
    autoskip=yes
    @18 "TO END PROGRAM" color=red
  #31 @75 "GFVB/JDH" attr=rev_video color=red;
  %DISPLAY WELCOME blank delete;
```

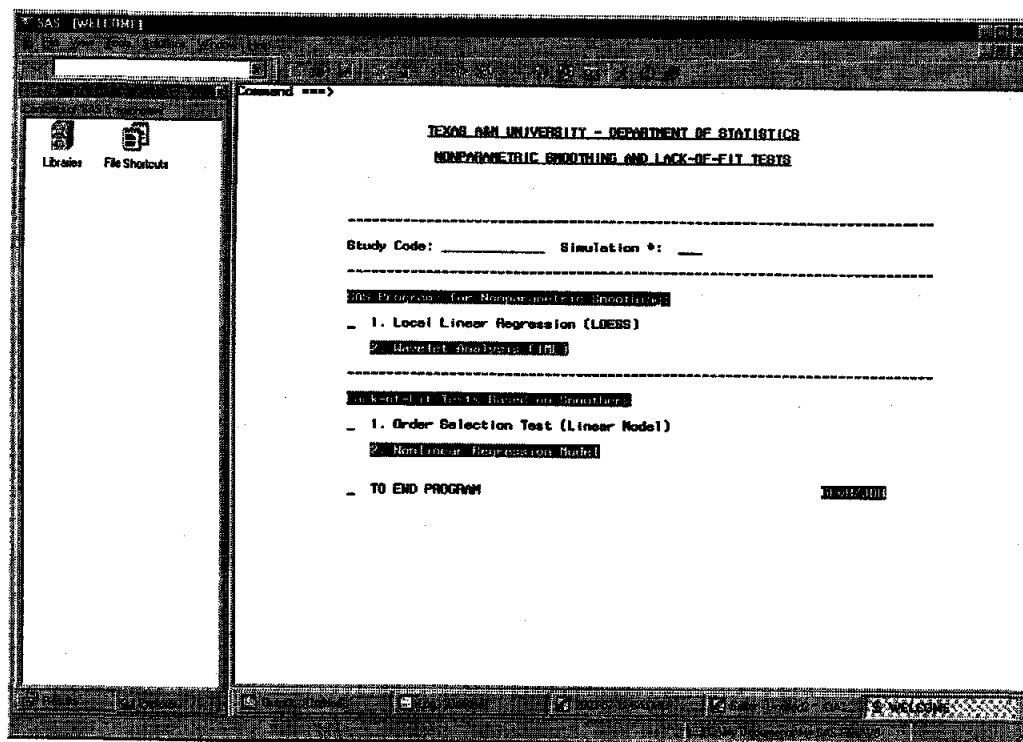


Figure 1: Welcome Window

The functions used in *%Window* are presented in detail in SAS Guide to Macro Processing [1] and Mace [7]. This command can be implemented in a data set, as we did in previous code, but the user has to submit the program every time he wants to display the window. A very useful way to eliminate this repetitive task is defining a macro. If for example we include our code in a macro definition as

```
%macro TESTE;
... command lines...
%mend TESTE;
```

then after submitting the program for the first time we can invoke it again only using the command *%TESTE*¹.

Another useful tool to eliminate repetitive tasks is the *Macro Autocall Facility*. It is used with macros to avoid submitting the program every time a new SAS session is started. The following piece of code gives the *Macro Autocall Facility* commands.

```
libname scsug02 'c:\sascode';
options mstored sasmstore=scsug02;
%macro TESTE / store des='SCSUG 2002';
... command lines...
%mend TESTE;
```

The new commands have the following functions:

- *libname scsug02 'c:\sascode';*
creates library *scsug02* on directory *c:\sascode*

¹You can try it with our code.

²In SAS Explorer select libraries and *scsug02*. Now scroll the cursor to the right to find the column description.

³The number *X/Y* indicates that the program has *Y* SAS tools for nonparametric smoothing and *X* lack-of-fit tests implemented

⁴Webpage <http://stat.tamu.edu/~gborries>

- *options mstore sasmstore=scsug02;*
says that a macro will be stored in the library indicated by *sasmstore*.
- */ store des='SCSUG 2002';*
writes macro *TESTE* in library indicated by *sasmstore*. Note that a file called "sasmacr" is now saved in the directory *c:\sascode*. The last part of the program creates a description for macro *TESTE*².

Details about these and many other commands using SAS Macros can be found in [1] and [4].

SASMLF is an example of application using macro language. It is composed of a group of macros for nonparametric smoothing and lack-of-fit ratio test in a user friendly way. In SASMLF 2/3 ³ we implemented

A SAS Programs for Nonparametric Smoothing

- 1 Proc LOESS for Local Linear Regression
- 2 Proc KDE for Kernel Density Estimation
- 3 Proc Univariate option for Histogram with Kernel Estimate

B Lack-of-Fit Tests Based on Smoothers

- 1 Order Selection Test for Linear Models
- 2 Ratio of Quadratic Forms for Linear Models

Other options being implemented are highlighted in Figure 2.

The setup file for SASMLF macro is available on Internet for download⁴. A series of movies with instructions about how to use the program and its applications are also available.

```

Command ===)

          TEXAS A&M UNIVERSITY - DEPARTMENT OF STATISTICS
          NONPARAMETRIC SMOOTHING AND LACK-OF-FIT TESTS

-----
Study Code: _____ Simulation #: _____ Catalog Name on sasuser: _____
                                     (Default is sasuser.smooth)
-----

SAS Programs for Nonparametric Smoothing:
- 1. Local Linear Regression (LOESS)
- 2. Kernel Density Estimation (KDE)
- 3. Histogram with Kernel Estimate (UNIVARIATE)
- 4. Related Functions (LANDAU)
- 5. Generalized Additive Models (GAM)
- 6. Thin-plate Smoothing Splines (TPSPLINE)
- 7. Wavelet Analysis (WAVE)

-----

Lack-of-Fit Tests Based on Smoothers:
- 1. Order Selection Test (Linear Model)
- 2. Ratio of Quadratic Forms (Linear Model)
- 3. Order Selection Test II (GLM)
- 4. Probability Bands (Linear Model)
- 5. Nonlinear Regression Model
- 6. Bayes Order Selection Test (Linear)

-----

Output Options:
- PDF - HTML - RTF          Save in _____

-----

- TO END PROGRAM                                [F10]

```

Figure 2: Welcome Window - SASMLF Macro

5 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 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] *SAS Guide to Macro Processing*. Cary, NC: SAS Institute Inc., second edition, 1994. version 6.
- [2] *SAS Macro Facility Tips and Techniques*. Cary, NC: SAS Institute Inc., second edition, 1994. version 6.
- [3] *SAS Guide to Applications Development, User's Guide*. Cary, NC: SAS Institute Inc., 1999.

- [4] Carpenter, A. *Carpenter's Complete Guide to the SAS Macro Language*. Cary, NC: SAS Institute Inc., 1998.
- [5] Hall, P. *The Bootstrap and Edgeworth Expansion*. Springer-Verlag, 1992.
- [6] Hart, J.D. *Nonparametric Smoothing and Lack-of-Fit Tests*. Springer-Verlag, 1997.
- [7] Mace, M.A. Orlando, Florida, 2002. SUGI 27. SAS User's Group International.

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

George F. von Borries, Ph.D. Student

Office: 506 Blocker Building
College Station - TX
77843 USA
Phone: (979) 845-3141
Fax: (979) 845-3144
E-Mail: gborries@stat.tamu.edu
Web: stat.tamu.edu/~gborries

Jeffrey D. Hart, Professor

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 Brasilia, 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*.

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

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are registered trademarks or trademarks of their respective companies.