PHILADELPHIA SAS USERS GROUP FALL 2014 MEETING

POSTER PRESENTATION

FUNCTIONAL LOGISTIC REGRESSION IN SAS

LINLIN FAN
E-MAIL: linlin.lynn.fan@gmail.com

DEPARTMENT OF MATHEMATICS

LEHIGH UNIVERSITY

OCTOBER 2014

1 ABSTRACT

While Functional Regression (Ramsay and Dalzell 1991) enables the task for regressing a scalar response on an infinite-dimensional (functional) predictor, modeling dichotomous response with functional predictor calls for special treatment. Functional Logistic Regression is part of the Functional Generalized Linear Models (James, 2002; Müller and Stadtmüller, 2005) which are in the framework of Functional Data Analysis (FDA; Ramsay and Silverman 1997) approaches. Functional Generalized Linear Models are evaluating the dynamic association (through $\beta(t)$ and appropriate link function $g(\cdot)$) between functional predictor ($X(t)$) and response (Y). Similar to Logistic Regression, Functional Logistic Regression facilitates a model-based classification of high-dimensional and low sample size data.

$$g(\mu) = \beta_0 + \int_T \beta(t)X(t)dt = \beta_0 + \sum_{k=1}^{\infty} \beta_k \xi_k \,, \mu = \mathrm{E}(Y)$$

Here, $\beta(t)$ and $X(t)$ are assumed to be spanned by the same eigen-basis. ξ_k is functional principal component score and $\varphi_k(t)$ is eigenfunction.

To configure the functional data, Functional Principal Component Analysis (FPCA; Rice and Silverman 1991) was first adopted in this work. The implementation of Functional Logistic Regression was done by SAS/IML. Several IML modules were written to build the whole pipeline of analysis.

2 CANADIAN WEATHER DATA INTRODUCTION

CanadianWeather dataset (Ramsay and Silverman 1997) was investigated for the classification of the categories of annual precipitation based on temperature profiles of a set of weather observatories across Canada. Annual precipitations for the 35 observatories across Canada are classified as two categories which will be detailed described in section 5. In Figure 1, there are 35 curves which represent 35 temperature profiles for the 35 observatories across Canada.

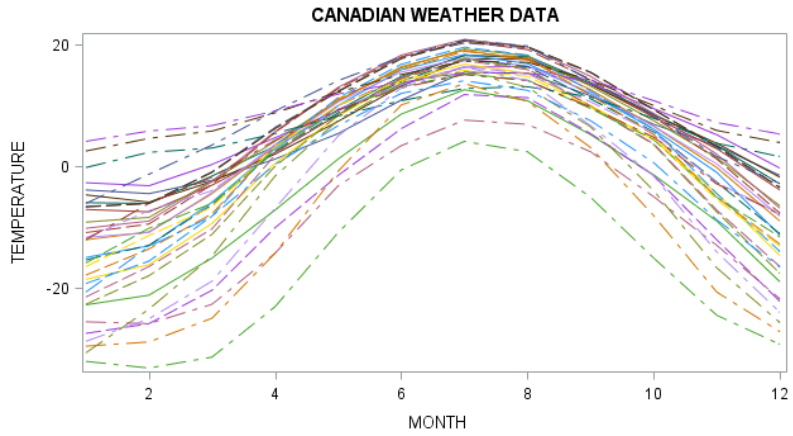
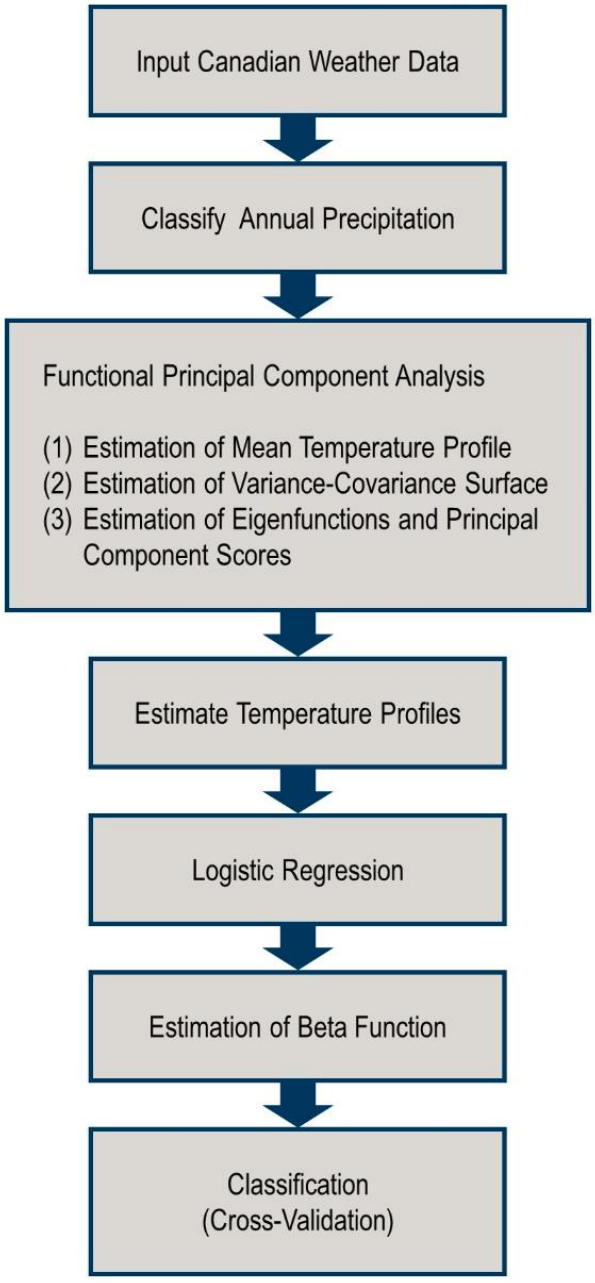


Figure 1: Temperature Profiles Plot

3 ASSUMPTIONS AND METHODOLOGIES

1) Assumptions
   (1) $\beta(t)$ and $X(t)$ are assumed to be spanned by the same eigen-basis.
   (2) ε are uncorrelated.
2) Methodologies
   (1) Functional Principal Component Analysis (FPCA)
   (2) Functional Logistic Regression

4 IMPLEMENTATION PIPELINE



5 CLASSIFY ANNUAL PRECIPITATIONS

There are two categories for annual precipitation. One is larger than or equal to the median of annual precipitations of the 35 observatories across Canada. The other one is less than the median of annual precipitations of the 35 observatories across Canada. The first one is labeled as 1, and the second one is labeled as 0.

6 ESTIMATION OF MEAN TEMPERATURE PROFILE

Mean temperature profile is estimated through conducting Non-Parametric Regression by using Local Linear Fit with different choices of weighting function.

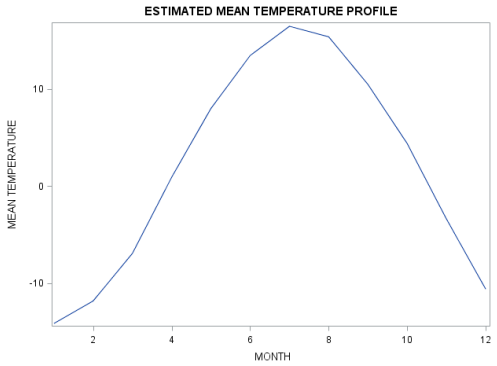


Figure 2: Estimated Mean Temperature Profile Plot

7 ESTIMATION OF SMOOTHED VARIANCE-COVARIANCE SURFACE

For estimating Variance-Covariance Surface, the following equations are applied. Since the estimated Variance-Covariance Surface is quite bumpy, Multiple Weighted Least Square is applied to smooth the surface.

$$\hat{\Sigma} = \frac{1}{35}\sum_{i=1}^{35} x_i^*(t)^T x_i^*(t) \ \ where\ x_i^*(t) = x_i(t) - \hat{\mu}(t)$$

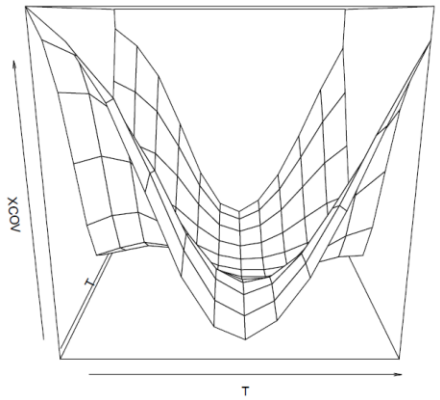


Figure 3: Estimated Smoothed Variance-Covariance Surface Plot

8 CHOICE OF COMPONENTS

Based on the estimated smoothed Variance-Covariance Surface, its eigenpairs can be obtained. Through the following equation, Scree Plot is generated in Figure 4. The first three components explain more than 99% of total variance.

$$FTV = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{12} \lambda_i}, k = 1, \cdots, 12$$

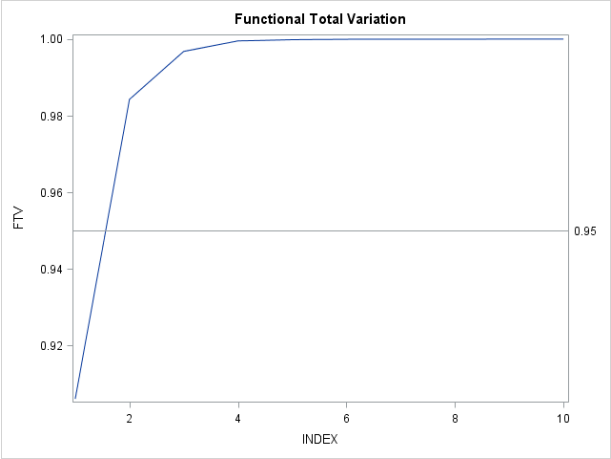


Figure 4: Scree Plot

9 ESTIMATION OF EIGENFUNCTIONS AND PRINCIPAL COMPONENT SCORES

For estimating eigenfunctions, based on properties of eigenpairs, the following equation needs to be satisfied.

$$\int_T \varphi_k(t)^T \varphi_k(t)\, dt = 1 \; where \; k = 1, 2, 3$$

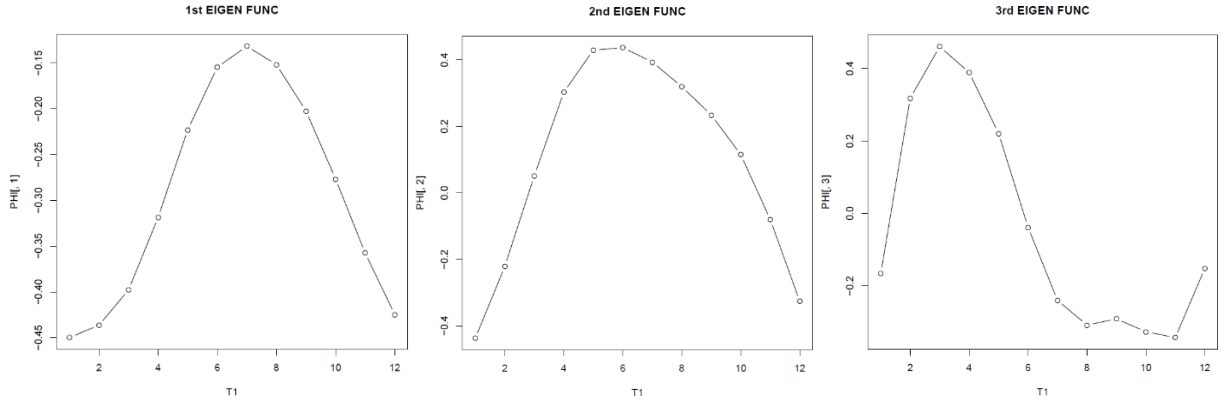


Figure 5: First Three Eigenfunctions Plot

Principal Component Scores can be approximated based on the following equation.

$$\xi_{ik} = \int_T [x_i(t) - \mu(t)]^T \varphi_k(t)\, dt \; where \; i = 1, 2, \cdots, 35 \; and \; k = 1, 2, 3$$

10 ESTIMATION OF TEMPERATURE PROFILES

Based on the following equation, the predicted temperature profiles are obtained. In Figure 6, dots represent true values and lines represent estimated temperature profiles.

$$\hat{x}_i(t) = \hat{\mu}(t) + \sum_{k=1}^{3} \hat{\xi}_{ik} \hat{\varphi}_k(t)^T \; where \; i = 1, 2, \cdots, 35$$

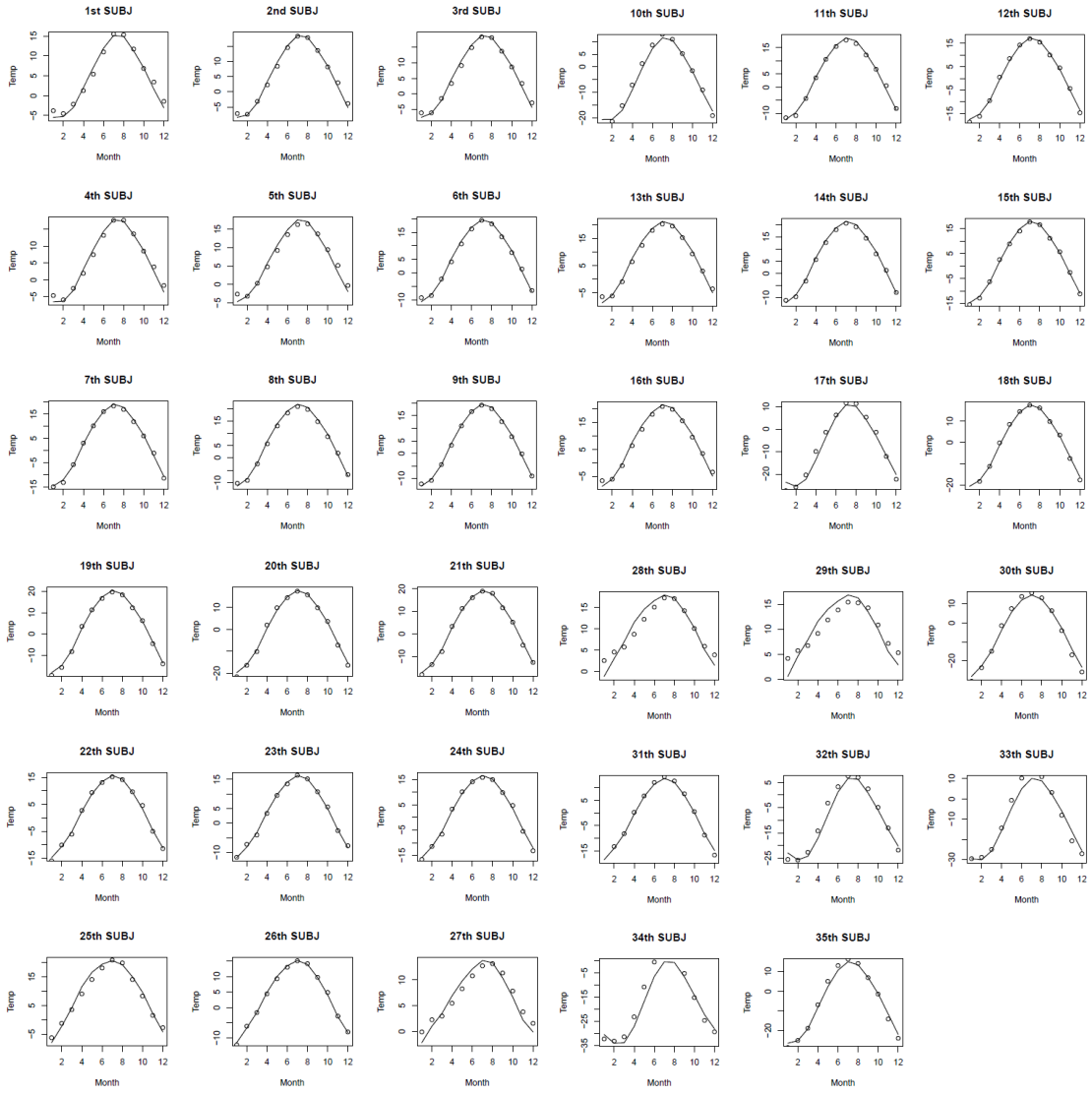


Figure 6: Predicted Temperature Profiles Plot

11 MODEL SELCTION

Akaike Information Criterion (AIC) is applied to make further choice of components. Through minimizing –AIC, the first three components are included in the model.

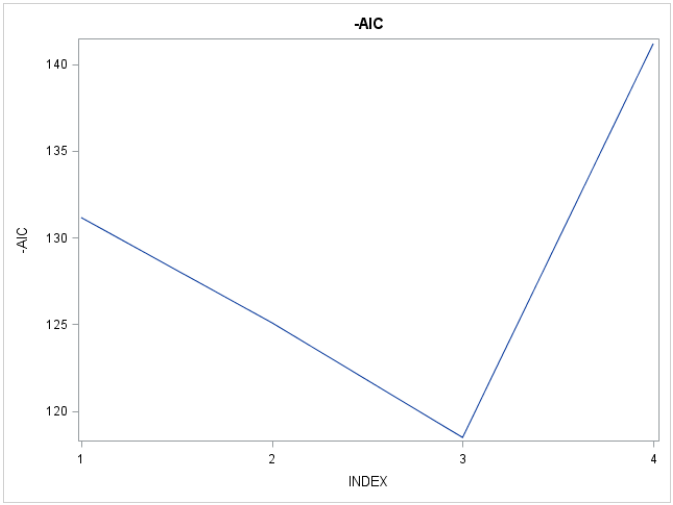


Figure 7: -AIC Plot

12 ESTIMATION OF BETA FUNCTION

Logistic Regression is applied to estimate β_k. Through the following equation, Beta Function is estimated.

$$\hat{\beta}(t) = \sum_{k=1}^{3} \hat{\beta}_k \hat{\varphi}_k(t)$$

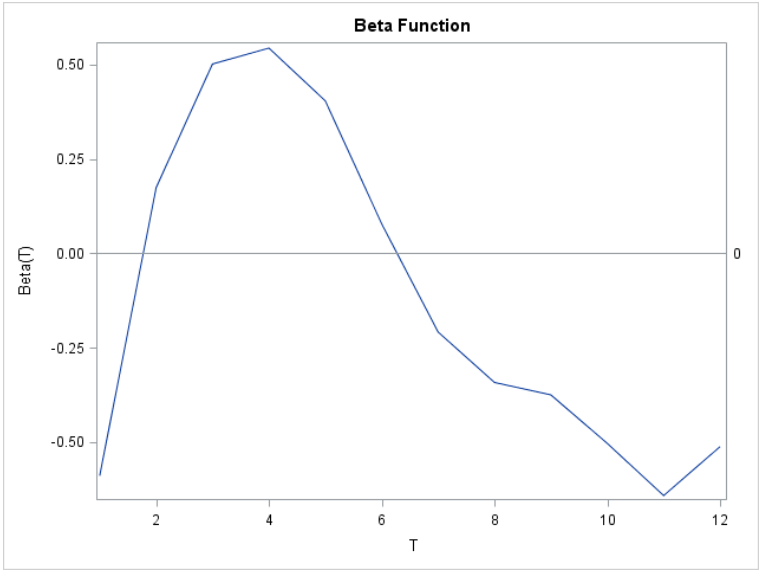


Figure 8: Beta Function Plot

13 VALIDATION

Based on the estimated beta function, the category of the annual precipitation for each observatory can be estimated. Through applying leave one out cross-validation, the misclassification rate is 0.1 by using Functional Logistic Regression. Therefore, in terms of misclassification rate, Functional Logistic Regression performs well.

When we apply Functional Logistic Regression, the AIC value for assessing goodness of fit of a model is 22.0134. When we apply Logistic Regression, the AIC value for assessing goodness of fit of a model is 26. Smaller AIC value is preferred, therefore, in terms of model fitting, Functional Logistic Regression is better than Logistic Regression.

14 CONCLUSION

The purposes of this analysis are to investigate the classification of the categories of annual precipitation (high, low) and get an insight of which period of time yields more classification power for annual precipitation through the implementation of Functional Logistic Regression with SAS/IML and IML Plus. Firstly, FPCA is applied to return first three principal components which leads infinite-dimensional functional data problem converts to finite-dimensional problem. In Figure 6, it turns out that the reconstruction of functional data performs well. Secondly, Beta Function is estimated by conducting Functional Logistic Regression. Through Figure 8, we can see that January, March, April, May, November, and December have more classification power. Thirdly, the misclassification rate for estimated response is 0.1 which indicates that Functional Logistic Regression performs well.

15 REFERENCE

This project is under the supervision of Professor Ping-Shi Wu.

[1] Rice, J., Silverman, B. (1991). "Estimating the Mean and Covariance Structure Nonparametrically When the Data are Curves". *J. Royal Statist. Soc. B* 53 (1): 233–243.

[2] Ramsay, J.O. and Dalzell, C.J. (1991). Some tools for functional data analysis. *J. Royal Statist. Soc. B* 53, 539-572.

[3] Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis*, Springer, New York.

[4] James, G.M. (2002). "Generalized Linear Models with Functional Predictors". Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64: 411–432.

[5] Müller, H.G. and Stadtmüller, U. (2005). "Generalized Functional Linear Models". The Annals of Statistics, **33**(2), 774–805.

```
/*************************************************/
/* PHILADELPHIA SAS USERS GROUP FALL 2014 MEETING */
/* FUNCTIONAL LOGISTIC REGRESSION IN SAS          */
/* SAS/IML MODULES                                */
/*************************************************/
PROC IML;
  RESET STORAGE=M450FDA.MYSTOR;
  START MULLWLSK(X,Y,GRID,H,KERNEL);
    ID1=LOC(X[1,] >= GRID[,1]-H[,1] && X[1,] <= GRID[,1]+H[,1]);
    ID2=LOC(X[2,] >= GRID[,2]-H[,2] && X[2,] <= GRID[,2]+H[,2]);
    ID=XSECT(ID1,ID2);
    LX=X[,ID];
    LY=Y[,ID];
    IN=NCOL(LX);
    PI=CONSTANT('PI');
    AX=(T(LX[1,])-GRID[,1])||(T(LX[2,])-GRID[,2]);
    IF KERNEL=1 THEN DO;
      U1=(LX[1,]-GRID[,1])/H[,1];
      U2=(LX[2,]-GRID[,2])/H[,2];
      W=DIAG(0.75*(1-U1##2))#DIAG(0.75*(1-U2##2));
    END;
    ELSE IF KERNEL=2 THEN DO;
      U1=(LX[1,]-GRID[,1])/H[,1];
      U2=(LX[2,]-GRID[,2])/H[,2];
      W=DIAG((1/((2*PI)**0.5))*EXP(-0.5*U1##2)#(1/((2*PI)**0.5))*EXP(-
0.5*U2##2));
    END;
    ELSE IF KERNEL=3 THEN DO;
      U1=(LX[1,]-GRID[,1])/H[,1];
      U2=(LX[2,]-GRID[,2])/H[,2];
      W=DIAG((15/16)*(1-U1##2))#DIAG((15/16)*(1-U2##2));
    END;
    ELSE W=DIAG(J(lN,1)/(2**2));
    FIT=WLS(AX,LY`,W);
    YHAT=FIT[1,];
    RETURN (YHAT);
  FINISH;

  START WLS(X,Y,W);
    X=J(NROW(X),1)||X;
    XPX=T(X)*W*X;
    XPY=T(X)*W*Y;

    BWLS=GINV(XPX)*(XPY);
    RETURN(BWLS);
  FINISH;

  START TRAPZINT(X,Y);
    AREA=SUM((Y[2:NROW(Y)]+Y[1:(NROW(Y)-1)])/2#(X[2:NROW(X)]-X[1:(NROW(X)-
1)]));
    RETURN(AREA);
  FINISH;

  START NPRegLLF(KERNEL,KNN,HK,X,Y,T);
    IF (KERNEL=1) | (KERNEL=2) | (KERNEL=3) | (KERNEL=4) THEN DO;
      IF KERNEL=2 THEN A=3; ELSE A=1;
    END;
```

```
    ELSE DO;
      PRINT 'The specification of kernel is unrecognized, Ep Kerenl is
adopted.';
      KERNEL=1;
    END;
    IF (KNN^=1 & HK <= 0) THEN DO;
      PRINT 'Invalid bandwidth is specifized, ROT is adopted.';
      HK=2.15*NROW(X)**(-0.2)*STD(X);
    END;
    IF (KNN=1 & HK < 0) THEN DO;
      PRINT 'Invalid K is specifized, 10 is adopted.';
      HK=10;
    END;
    IF KNN=1 THEN K=HK;
    YHAT=J(12,1,.);
    PI=CONSTANT('PI');
    DO i=1 TO 12;
      IF KNN=1 THEN DO;
        DIST=ABS(X-T[i]);
        CALL SORTNDX(DX,DIST,1);
        HK=DIST[DX[K,],];
        IDX=DX[1:K];
      IF KERNEL=2 THEN HK=HK/3;
    END;
    ELSE IDX=LOC(ABS((X-T[i])/(A*HK))<=1);
    U=(X[IDX]-T[i])/HK;
    IF KERNEL=1 THEN W=DIAG(0.75*(1-U##2));
    ELSE IF KERNEL=2 THEN W=DIAG((1/((2*PI)**0.5))*EXP(-(U##2)/2));
    ELSE IF KERNEL=3 THEN W=DIAG((15/16)*(1-U##2)##2);
    ELSE W=DIAG(J(NROW(X[IDX]),1)/2);
    FIT=WLS((X[IDX]-T[i]),Y[IDX],W);
    YHAT[i]=FIT[1,];
    END;
    RETURN (YHAT);
  FINISH;

  START GETMU(KERNEL,KNN,HK,X,Y,T,N);
    IF NROW(Y)=N THEN VY=COLVEC(Y);
      ELSE VY=COLVEC(T(Y));
      IF NROW(X)=N THEN VX=COLVEC(X);
      ELSE VX=COLVEC(T(X));
    MUHAT=NPRegLLF(KERNEL,KNN,HK,VX,VY,T);
    RETURN (MUHAT);
  FINISH;

  START GETXCOV(T,Y,MUHAT,NTOUT,H,KERNEL,XCOV,TGRID,SQXCOV);
    YSTAR=Y-REPEAT(MUHAT,1,NCOL(Y));
    SIGMA=YSTAR*YSTAR`/NCOL(Y);
    CXXN=COLVEC(SIGMA`);
    XX1=REPEAT(T,1,12);
    XX1=COLVEC(XX1);
    XX2=REPEAT(T,1,12);
    XX2=COLVEC(XX2`);
    INC=(T[<>,]-T[><,])/(NTOUT-1);
    TGRID=T(DO(T[><,],T[<>,],INC));
    TT1=REPEAT(TGRID`,NTOUT,1);
    TT1=COLVEC(TT1`);
```

```
    TT2=REPEAT(TGRID,1,NTOUT);
    TT2=COLVEC(TT2`);
    TMESH=TT1||TT2;
    NTT=NROW(TMESH);
    TPAIR=T(XX1||XX2);
    YY=J(NTT,1);
    DO I=1 TO NTT;
      YY[I,]=MULLWLSK(TPAIR,CXXN`,TMESH[I,],H,KERNEL);
    END;
    XCOV=TMESH||YY;
    SCOV=I(12);
    DO I=1 to 12;
      DO J=1 to 12;
        SCOV[I,J]=YY[(I-1)*12+J,];
      END;
    END;
    SQXCOV=(SCOV+SCOV`)/2;
  FINISH;

  START GETPCS(T,Y,MUHAT,TGRID,SQXCOV,M,EVEC,EVAL,PHI,XIHAT,YPRED,XCOV_EST);
    INC=(TGRID[<>,]-TGRID[><,])/(NROW(TGRID)-1);
    CALL EIGEN(EVAL,EVEC,SQXCOV);
    EVAL=EVAL[1:M,]*INC;
    EVEC=EVEC[,1:M]/SQRT(INC);
    PHI=J(NROW(T),M);
    DO I=1 TO M;
      EVEC[,I]=EVEC[,I]/SQRT(TRAPZINT(TGRID,EVEC[,I]##2));
        IF EVEC[2,I] < EVEC[1,I] THEN EVEC[,I]=-EVEC[,I];
      TMPDATA=TGRID||EVEC[,I];
      CALL SPLINEC(TMP,COEFF,ESLOPES,TMPDATA);
      YTMP=SPLINEV(COEFF,T);
      YTMP[NROW(YTMP),]=TMP[NROW(TMP),];
      PHI[,I]=YTMP[,2];
    END;
    N=NCOL(Y);
    YY=Y-REPEAT(MUHAT,1,N);
    XIHAT=J(N,M);
    DO I=1 TO N;
      DO J=1 TO M;
        PROD=YY[,I]#PHI[,J];
        XIHAT[I,J]=TRAPZINT(T,PROD);
        END;
    END;
    YPRED=REPEAT(MUHAT,1,N)+PHI*XIHAT`;
    XCOV_EST=PHI*DIAG(EVAL)*PHI`;
  FINISH;
  STORE MODULE=(WLS NPRegLLF GETMU MULLWLSK GETXCOV GETPCS TRAPZINT);
QUIT;
```