III Jornadas de usuarios de R

17 y 18 de Noviembre de 2011
Escuela de Organización Industrial, Madrid

Modelización y Predicción con Datos Funcionales en R

Grupo de Investigación: Modelización y Predicción con Datos Funcionales

Departamento de Estadística e I.O, Universidad de Granada

Objetivos

- Dar a conocer nuestro trabajo, así como las recientes aportaciones sobre modelización y predicción de Datos Funcionales, haciendo uso de R

- Planteamiento de problemas y espera de sugerencias

Introducción al FDA

- **Dato funcional**: en general, es una curva que procede de la realización de un proceso estocástico en tiempo continuo
- **Muestra**: conjunto de funciones observadas en instantes de tiempo
- El argumento de las curvas no siempre es el tiempo (Quimiometría)

Ejemplos reales y aplicaciones con datos funcionales

- Curvas de temperaturas para diferentes zonas geográficas de Andalucía

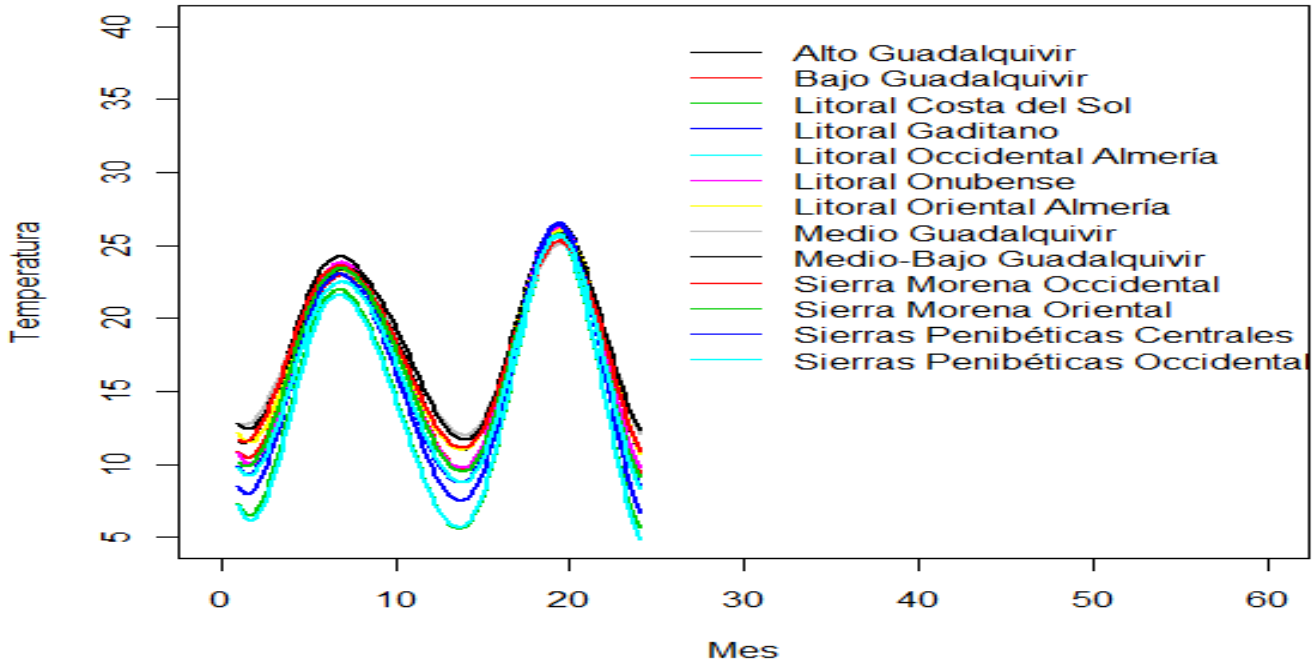


Figure: Curvas de temperatura para 13 zonas geográficas de Andalucía

Ejemplos reales y aplicaciones con datos funcionales

- **Predicción de las cotizaciones en bolsa de Madrid del grupo banca**
  - La evolución temporal de las cotizaciones bursátiles se puede modelizar como una realización de un proceso estocástico en tiempo continuo.
  - Aguilera et al. (1999) propusieron un modelo de predicción en componentes principales (PCP model) para predecir la curva de cotizaciones bursátiles en la primeras cinco semanas de 1997 (futuro) a partir de su evolución temporal desde 1992 (pasado).

Ejemplos reales y aplicaciones con datos funcionales

- **Modelización de la relación entre estrés y el lupus**
  - El Lupus Eritomatoso es una enfermedad autoinmune altamente relacionada con el nivel de estrés al que está sometido el individuo.
  - Muestra: 44 pacientes de lupus para los que se disponía de observaciones diarias de su nivel de estrés durante un periodo de 18 días.
  - Se define una variable respuesta binaria con valor uno para pacientes con brote y valor cero para el resto.
  - Modelo de Regresión Logística Funcional en base al ACPF de las curvas muestrales (Aguilera et al., 2008).

Ejemplos reales y aplicaciones con datos funcionales

- **Predicción de las curvas de concentración de polen a partir de las curvas de temperatura**
  - Muestra: observaciones diarias tomadas en el Departamento de Aerobiología de la Facultad de Ciencias de la Universidad de Granada durante los últimos años (hasta 2009).
  - Modelo Lineal Funcional de Respuesta Funcional (Valderrama et al., 2009).

Obtención de la forma funcional

En la práctica se dispone de **observaciones discretas**, de modo que el primer paso en ADF es reconstruir la **forma funcional**

- Información muestral $\rightarrow x_1(t), x_2(t), \ldots, x_n(t)$
- Variable funcional $\rightarrow X = \{X(t) : t \in T\}$
- Funciones muestrales pertenecientes a $L^2(T)$

Obtención de la forma funcional

En la práctica se dispone de **observaciones discretas**, de modo que el primer paso en ADF es reconstruir la **forma funcional**

- Información muestral $\rightarrow x_1(t), x_2(t), \ldots, x_n(t)$
- Variable funcional $\rightarrow X = \{X(t) : t \in T\}$
- Funciones muestrales pertenecientes a $L^2(T)$

Problema: Observaciones discretas, x_{ik}, de $x_i(t)$ en $\{t_{i0}, t_{i1}, \ldots, t_{im_i} \in T,\ i = 1, \ldots, n\}$

Obtención de la forma funcional

En la práctica se dispone de **observaciones discretas**, de modo que el primer paso en ADF es reconstruir la **forma funcional**

- Información muestral $\rightarrow x_1(t), x_2(t), \ldots, x_n(t)$
- Variable funcional $\rightarrow X = \{X(t) : t \in T\}$
- Funciones muestrales pertenecientes a $L^2(T)$

Problema: Observaciones discretas, x_{ik}, de $x_i(t)$ en $\{t_{i0}, t_{i1}, \ldots, t_{im_i} \in T,\ i = 1, \ldots, n\}$

Solución: Representación básica

- Base: $\{\phi_1(t), \ldots, \phi_p(t)\}$

Representación básica: $x_i(t) = \sum_{j=1}^{p} a_{ij}\phi_j(t),\ i = 1, \ldots, n$

Bases usuales

- Bases de B-splines

Bases usuales

- Bases de B-splines
- Sistemas base de Fourier

Bases usuales

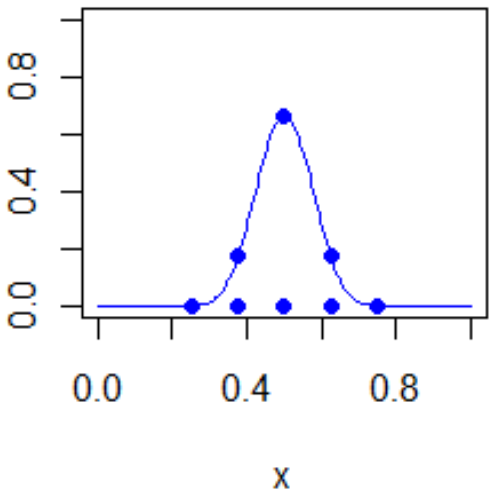
- Bases de B-splines
- Sistemas base de Fourier
- Bases de Wavelets

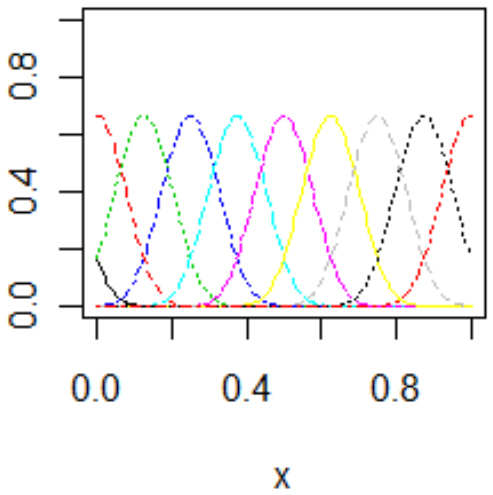
Bases usuales

- Bases de B-splines
- Sistemas base de Fourier
- Bases de Wavelets
- Bases de potencias y exponenciales
- Base de funciones constantes
- Bases polinomiales

Bases usuales

- B-splines Cúbicos: $B_{j,4}(t)$ (De Boor, 2002)





Modelización y Predicción de Datos Funcionales

Problema 1

Gran dimensión de los datos y Multicolinealidad

Solución 1

Análisis en Componentes Principales Funcional

Modelización y Predicción de Datos Funcionales

Problema 1

Gran dimensión de los datos y Multicolinealidad

Solución 1

Análisis en Componentes Principales Funcional

Problema 2

Los datos son observados con error o ruido

Solución 2

Penalización de la rugosidad (Roughness Penalty)

Recursos disponibles en R

- Librería "fda" (J. O. Ramsay, Hadley Wickham, Spencer Graves, Giles Hooker)
- Librería "fda.usc" (M. Febrero-Bande y M. Oviedo de la Fuente) (No paramétrica)

Principales aportaciones en R: Penalización

- ACP Funcional $\equiv$ ACP multivariante de la matriz $A\Psi^{1/2}$ (Ocaña et al., 2007)
  - $A = (a_{ij})_{nxp}$ matriz de coeficientes básicos de las trayectorias muestrales
  - $\Psi = (\Psi_{ij})_{p\ x\ p} = \int \phi_i(t)\,\phi_j(t)\,dt$ matriz del producto interior entre dos funciones base.

Principales aportaciones en R: Penalización

- ACP Funcional $\equiv$ ACP multivariante de la matriz $A\Psi^{1/2}$ (Ocaña et al., 2007)
  - $A = (a_{ij})_{nxp}$ matriz de coeficientes básicos de las trayectorias muestrales
  - $\Psi = (\Psi_{ij})_{p\ x\ p} = \int \phi_i(t)\,\phi_j(t)\,dt$ matriz del producto interior entre dos funciones base.

- ACP Funcional Penalizado Tipo I
  1. Suavización de las trayectorias muestrales usando Penalized splines (P-splines)
  2. ACP Funcional de los P-splines

- ACP Funcional Penalizado Tipo II
  - Se introduce la penalización en la contrucción de las CPs.

Principales aportaciones en R: Penalización

- Regresión Logit Funcional en Componentes Principales
  - Opción I: covariables = CPs del ACPF tipo I
  - Opción II: covariables = CPs del ACPF tipo II

- Regresión Logit Funcional Penalizada
  - Penalización en el criterio de Máxima verosimilitud como técnica de reducción de la dimensión

Trabajando con R

- Library(fda), library(splines), library(zoo), library(glm), library(lrm),...
- Base de B-splines Cúbicos: create.bspline.basis(...)
- Penalized splines (P-splines)

```
#----------------------Curvas ajustadas con P-splines-------------------------
    #-------âi=(psi'psi+lambd p'p)^-1 * psi'data[,i]---------------------------
    CurvasPsplines<-function(datos,nnodos,lambda,tiempos,tiemposNew,objbase,n){
        p=diff(diff(diag(nnodos+2)))
        penmatriz<-t(p) %*% p
        Acoef <- matrix(0, ncol=n,nrow=nnodos+2)
        evalbase<-eval.basis(tiempos,objbase)
        inversa<-solve((t(evalbase)%*% evalbase) + (lambda*penmatriz))
        for (i in 1:n){Acoef[,i]<- inversa%*% t(evalbase)%*% datos[,i]}
        #Curvas
        evalbase2<-eval.basis(tiemposNew,objbase)
        curvasPspl <- evalbase2%*%Acoef
        return(curvasPspl,Acoef)}
```

Trabajando con R: Procedimientos para la selección de parámetros

```
GCV_PCLR_SPCA <- function(datosTr,yTr,lambda,nbasis,objbase,tiempos,tiempos2,argnames,n,m,m2,nnodos){
    GCV <- c()
      for (p in 1:nbasis){
          SPCA <- spca(datosTr,nnodos,tiempos,tiempos2,objbase,lambda,n,m,m2)
          X <- SPCA$scoresSPCA[,1:p]
          modeloLogit <- glm(yTr~. ,family=binomial(link="logit"), data=X)
          y_estim <- predict(modeloLogit,newdata=X,type="response")
          W <- matrix(0,length(y_estim),length(y_estim))
          sumas <- 0
          ...
          ECM <- sumas
          HAT <- sqrtW%*%X%*%solve(t(X)%*%W%*%X)%*%t(X)%*%sqrtW
          trazaH <- sum(diag(HAT))
          GCV[p] <- ECM/(n-trazaH)^2
      }
      num_cp <- 0
      k <- 2
      while(k<= nbasis-1){
        if(GCV[k]<GCV[k-1] && GCV[k]<GCV[k+1]){
          num_cp <- k
          break
          }else{k=k+1}
      }
      GCVerror<- GCV[k]
return(num_cp, lambda,GCV,GCVerror)}
```

Problemas computacionales de la Penalización

PENALIZACIÓN: Matriz de Penalización + Parámetro de suavizado (λ)

Es fundamental la selección del parámetro de suavizado

Problemas computacionales de la Penalización

PENALIZACIÓN: Matriz de Penalización + Parámetro de suavizado (λ)

Es fundamental la selección del parámetro de suavizado

- Validación Cruzada Generalizada
- Validación Cruzada (leave-one-out)

Problemas computacionales de la Penalización

PENALIZACIÓN: Matriz de Penalización + Parámetro de suavizado (λ)

Es fundamental la selección del parámetro de suavizado

- Validación Cruzada Generalizada
- Validación Cruzada (leave-one-out)

$$\Downarrow$$

Elevado coste computacional (tiempo de cómputo)

Conclusiones y Problemas Abiertos

Conclusiones

- ADF mejora al Análisis multivariante
- Ante datos ruidoso, las técnicas penalizadas (para el ADF) aportan mejoras con respecto a las tradicionales o no penalizadas

Conclusiones y Problemas Abiertos

Conclusiones

- ADF mejora al Análisis multivariante
- Ante datos ruidoso, las técnicas penalizadas (para el ADF) aportan mejoras con respecto a las tradicionales o no penalizadas

Líneas abiertas

- Trabajo futuro: elaboración de una librería en R
  - ¿Cómo ahorrar en coste computacional?
  - ¿Es necesaria la implementación de las funciones en C++?

Referencias bibliográficas

- Aguilera A.M., Gutiérrez, R. and Valderrama, M.J. (1996). Approximation of estimators of the PCA of a stochastic process using B-splines. Communications in Statistics. Simulation and Computation, 25(3), 671-691.
- Aguilera, A.M., Ocaña, F.A. y Valderrama, M.J. (1999). Stochastic modelling for evolution of stock prices by means of functional PCA. Applied Stochastic Models in Business and Industry, 15(4), 227-234.
- Aguilera, A.M., Escabias, M. y Valderrama, M.J. (2008). Discussion of different logistic models with functional data. Application to Systemic Lupus Erythematosus. Computational Statistics and Data Analysis, 53(1), 151-163.
- De Boor, C. (1977). Package for calculating with B-splines. Journal of Numerical Analysis, 14, 441-472.
- Eilers, P. y Marx, B. (1996). Flexible smoothing with B-splines and penalties. Statistical Science, 11, 89-121.
- http://eio.usc.es/pub/MAESFE/ (librería fda.usc)
- http://www.functionaldata.org (librería fda)
- http://www.r-project.org/
- Ocaña, F.A., Aguilera, A.M. and Escabias, M. (2007). Computational considerations in functional principal component analysis. Computational Statistics, 22(3): 2145-2166.
- Ramsay, J. O. y Silverman, B. W. (1997, 2005). Functional data analysis (First and Second editions). Springer-Verlag.
- Valderrama, M.J., Ocaña, F.A. y Aguilera, A.M. (2009). Forecasting Pollen Concentration by a Two-Step Functional Model. Biometrics, en prensa (DOI: 10.1111/j.1541-0420.2009.01293.x).

GRACIAS POR SU ATENCIÓN.