

# Sensibilidad de procedimientos estadísticos: Bondad de Ajuste y Robustez.

**Alejandra Cabaña Nigro**

y

**Alfonso Gordaliza Ramos**

Departamento de Estadística e Investigación Operativa  
Universidad de Valladolid

- **José Ramón Berrendero** (Universidad Autónoma de Madrid), B. Méndez, D. Tyler y R. Zamar: *Curvas de sesgo para M-estimadores restringidos*.
- **Ricardo Cao** (Universidade da Coruña) e I. Van Keilegom: *Empirical Likelihood tests for two-sample problems via nonparametric density estimation*.
- **Antonio Cuevas** (Universidad Autónoma de Madrid), J.R. Berrendero y F. Vázquez : *Un test de uniformidad multivariante basado en una noción geométrica de profundidad*.
- **Pedro Delicado** (Universitat Politècnica de Catalunya) y J. Romo: *Constant coefficient tests for random coefficient regression*.
- **Luis Angel García Escudero** (Universidad de Valladolid) y A. Gordaliza : *Some applications for trimmed k-means clustering*.
- **Aurea Grané** (Universitat de Barcelona) y Josep Fortiana : *Estadísticos para contrastes de bondad de ajuste basados en correlaciones máximas*.
- **Wenceslao González Manteiga** (Universidad de Santiago de Compostela), J.C. Pardo e I. Van Keilegom: *Comparison of regression curves based on the estimation of the error distribution*.
- **Juan Romo** y S. López Pintado, (Universidad Carlos III): *Datos funcionales: la idea de profundidad*.

# Curvas de sesgo para M-estimadores restringidos

José R. Berrendero

Departamento de Matemáticas

Universidad Autónoma de Madrid

*Congreso MAT.ES, Valencia, febrero 2005*

Sesión especial:

*Sensibilidad de procedimientos estadísticos: Bondad de Ajuste y Robustez.*

Una medida muy informativa de la robustez de un estimador es su curva de sesgo máximo. Esta curva mide el máximo sesgo asintótico posible del estimador cuando se admite la presencia de una fracción  $\epsilon$  de datos contaminados en la muestra. En este trabajo se calculan y comparan las curvas de sesgo máximo de los MM-estimadores y de los M-estimadores restringidos de regresión. También se comparan las curvas de sesgo máximo de estos estimadores con las de otros que ya eran conocidas, como los S- y los  $\tau$ -estimadores. Los M-estimadores restringidos tienen el mejor comportamiento si se tienen en cuenta conjuntamente las propiedades de robustez –medidas a través de la curva de sesgo máximo– y de eficiencia bajo el modelo normal sin contaminación.

Otro resultado de interés es que para todo S-estimador existe un M-estimador restringido cuyo sesgo máximo nunca supera al del S-estimador y es estrictamente menor para alguna fracción de contaminación  $\epsilon$ . Es decir, que los S-estimadores –incluido el estimador que minimiza la mediana de los cuadrados de los residuos– son inadmisibles respecto al sesgo.

Este es un trabajo conjunto con **Beatriz Mendes, David Tyler y Rubén Zamar**.

# Empirical likelihood tests for two-sample problems via nonparametric density estimation.

Ricardo Cao and Ingrid Van Keilegom  
Universidade da Coruña and Université Catholique de Louvain.

We study the two-sample problem by comparing kernel estimators of the densities of the two populations via a local empirical likelihood approach. A bootstrap approximation is proposed to calibrate the empirical likelihood test. The proposed method can be easily extended to more than two samples and to multivariate distributions. Some simulation study illustrates the theory.

**Un test de uniformidad multivariante basado  
en una noción geométrica de profundidad**

Antonio Cuevas  
Departamento de Matemáticas  
Facultad de Ciencias  
Universidad Autónoma de Madrid

**Congreso MAT.ES, Valencia, Febrero de 2005**

Sesión especial dedicada a *Sensibilidad de procedimientos estadísticos:*  
*Bondad de Ajuste y Robustez*

Se propone y analiza un test de uniformidad multivariante para el caso en que la información disponible es una muestra aleatoria elegida en un recinto compacto  $S \subset \mathbb{R}^d$ .

Este test está basado en las distancias entre los puntos muestrales y la frontera de  $S$  y puede adaptarse al caso en que  $S$  es desconocido utilizando para ello técnicas de estimación de conjuntos.

El procedimiento es computacionalmente factible, incluso en dimensión superior a 2, y proporciona resultados competitivos cuando se compara con otros tests de uniformidad.

Este es un trabajo conjunto con **José Ramón Berrendero** y **Francisco Vázquez**, del Departamento de Matemáticas de la Universidad Autónoma de Madrid.

# CONSTANT COEFFICIENT TESTS FOR RANDOM COEFFICIENT REGRESSION.

Pedro Delicado and Juan Romo

Random coefficient regression models have been applied in different fields and they constitute a unifying setup for many statistical problems. The non-parametric study of this model started with Beran y Hall (1992) and it has become a fruitful framework. In this paper we propose and study statistics for testing a basic hypothesis concerning this model: the constancy of coefficients. The asymptotic behavior of the statistics is investigated and bootstrap approximations are used in order to determine the critical values of the test statistics. A simulation study illustrates the performance of the proposals.

# SOME APPLICATIONS FOR TRIMMED $k$ -MEANS CLUSTERING

L.A. García-Escudero and A. Gordaliza  
Departamento de Estadística e Investigación Operativa  
Universidad de Valladolid. Spain

Trimmed  $k$ -means methodology was introduced by Cuesta-Albertos, Gordaliza and Matrán (1997) with the aim of robustifying the  $k$ -means method (paradigm of the non-hierarchical cluster analysis) through the use of an “impartial” trimming procedure. The key idea is that data themselves tell us the observations to be trimmed off are.

In this talk, we will review the statement of the method and some basic properties illustrating the robustness gain. Some applications in the field of functional data clustering and clusterwise linear regression will be provided. Finally, a possible extension that allows the procedure to handle differently scattered groups will be presented.

## References

- [1] Cuesta-Albertos, J.A., Gordaliza, A. and Matrán, C. (1997), “Trimmed  $k$ -means: An attempt to robustify quantizers,” *The Annals of Statistics*, **25**, 553-576.

# Estadísticos para contrastes de bondad de ajuste basados en correlaciones máximas.

Aurea Grané, Josep Fortiana  
Departament d'Estadística, Facultat de Matemàtiques  
Universitat de Barcelona,  
agran@ub.edu, fortiana@ub.edu

Dada una muestra  $X_1, \dots, X_n$  iid  $\sim F$ , nos proponemos contrastar la hipótesis nula que  $F = F_0$ , donde  $F_0$  es una función de distribución completamente especificada.

En trabajos anteriores hemos estudiando la correlación máxima de Hoeffding,  $\rho^+$ , como una medida de proximidad entre la distribución empírica de los datos,  $F_n$  y  $F_0$ . Puesto que  $\rho^+$  presentaba varios inconvenientes como medida de bondad de ajuste, propusimos una leve modificación de la misma que dio lugar al estadístico  $Q_n$ . Este estadístico es proporcional a la diferencia media de Gini, es un  $L$ -estadístico, tiene mejores propiedades, es posible calcular su distribución exacta bajo la hipótesis nula y también su distribución asintótica bajo ambas hipótesis.  $Q_n$  tiene una descomposición en serie parecida a la de otros estadísticos de bondad de ajuste, como el de Cramér-von-Mises. Variando los coeficientes de esta descomposición se obtiene una clase de estadísticos, que son combinaciones lineales de las componentes de  $Q_n$ , con el objetivo de optimizar la potencia para una alternativa o familia de alternativas dadas.

Estos estadísticos así obtenidos no alcanzan en general la potencia óptima (Neyman-Pearson), pero mantienen las buenas propiedades de  $Q_n$ .

Las componentes de  $Q_n$  se obtienen a partir de las funciones propias del operador de covarianza de un proceso estocástico (asociadas a la descomposición de Karhunen-Love de dicho proceso).

Hemos visto que la base de Karhunen-Love no es óptima en general, puesto que, tal como ocurre con los *smooth tests* una base ortonormal será más adecuada para el test en la medida en que detecte bien la distribución alternativa.

La base de Karhunen-Love no tiene en cuenta la alternativa, pues se trata de una regresión funcional en componentes principales. Para incorporar la información sobre la distribución alternativa, construimos un estadístico de test a partir de una versión funcional de la regresión por mínimos cuadrados parciales (PLS).

*Palabras clave:*  $L$ -estadísticos, sucesiones ortogonales, bondad de ajuste, correlación máxima de Hoeffding.

# Comparison of regression curves based on the estimation of the error distribution

Juan Carlos Pardo-Fernández <sup>\*</sup>     Ingrid Van Keilegom <sup>†</sup>  
Wenceslao González-Manteiga <sup>‡</sup>

September 14, 2004

## ABSTRACT

Let  $(X_1, Y_1)$  and  $(X_2, Y_2)$  be two independent random vectors, and assume that they satisfy two non-parametric regression models  $Y_j = m_j(X_j) + \sigma_j(X_j)\varepsilon_j$ , for  $j = 1, 2$ , where the error variable  $\varepsilon_j$  is independent of  $X_j$ ,  $m_j(x) = E(Y_j|X_j = x)$  is an unknown regression function and  $\sigma_j^2(x) = \text{Var}(Y_j|X_j = x)$  is the conditional variance function. Suppose that  $\varepsilon_1$  and  $\varepsilon_2$  have common distribution with mean 0 and variance 1.

In this work we introduce a procedure to test the hypothesis  $H_0 : m_1(x) = m_2(x)$  of equality of the regression functions. The test is based on the comparison of the empirical distribution of the errors in each model (see Akritas and Van Keilegom, 2001). A Kolmogorov-Smirnov type statistic and a Cramér-von Mises type statistic are considered.

Some asymptotic results are proved: weak convergence of the process of interest, convergence of the test statistics. We describe a bootstrap procedure in order to get the critical values of the test. Finally, some simulations are presented to check the behaviour of this testing procedure with small and moderate sample sizes and one extension to the censored data case is given.

**Key words:** comparison of regression curves; goodness-of-fit; bootstrap.

---

<sup>\*</sup>Departamento de Estatística e I.O. Universidade de Vigo

<sup>†</sup>Institute de Statistique. Université Catholique de Louvain

<sup>‡</sup>Departamento de Estatística e I.O. Universidade de Santiago de Compostela



## References

- [1] Akritas, M. G. and Van Keilegom, I. (2001) *Nonparametric estimation of the residual distribution. Scand. J.Statist.*, 28, 549-568

## **Datos funcionales: la idea de profundidad**

*Sara López Pintado y Juan Romo*  
*Universidad Carlos III de Madrid*

La recopilación de observaciones en muchos ámbitos de la ciencia y la tecnología proporciona de manera natural datos que son funciones. La noción estadística de profundidad es robusta y de carácter no paramétrico y parece adecuada como herramienta para analizar observaciones funcionales. Se propone una definición de profundidad para funciones y se analizan sus propiedades. Además, la versión finito-dimensional de este nuevo concepto constituye una alternativa a los conceptos de profundidad ya existentes y es computacionalmente poco costoso en cualquier dimensión. Finalmente, se ilustra esta propuesta con diferentes ejemplos reales.