Goodness–of–Fit tests for regression models: the functional data case

Wenceslao González Manteiga
Joint work with Juan Cuesta–Albertos, Manuel Febrero–Bande and Eduardo García–Portugués

Department of Statistics and O.R.
University of Santiago de Compostela
Spain
Part I: Review of Goodness–of–Fit tests for regression models

1. Introduction
2. Calibration, size and power
3. Recent GOF tests for regression models
4. GOF in other contexts and some recent advances

Part II: The functional case: GOF of the functional linear model

1. Functional data
2. The test
3. Simulation study
4. Real data application
Recent collaborations with J. Cuesta–Albertos (UC), R.M. Crujeiras (USC), M. Febrero–Bande (USC), E. García–Portugués (USC), A. Monsalve (UCLA) and D. Rodríguez (UBA).
And many other collaborators in the last 20 years:
- C. Cadarso Suárez, L. A. Ramil Novo, C. Sánchez Sellero (USC)
- J. de Uña, J. C. Pardo Fernández, J. Roca Pardiñas, C. Campos Rodríguez, A. González-Perez (UV)
- T. Alcalá, J. A. Cristóbal (Universidad de Zaragoza), M. Delgado (Universidad Carlos III de Madrid), M. D. Martínez Miranda (Universidad de Granada)
- H. Koul (Michigan State University, USA), W. Stute (University of Giessen, Germany), I. Van Keilegom (Catholic University of Louvain, Belgium), P. Vieu (University Paul Sabatier, France), G. Boente (Universidad de Buenos Aires, Argentina)
Goodness of fit

The term *Goodness-of-fit* (GOF) was introduced by Pearson at the beginning of the 20th century and it refers to tests that check how a distribution fit to a data set in an omnibus way.

The basic idea consists in comparing a nonparametric pilot estimator for the unknown distribution $F$ or the density $f$, with a consistent parametric estimator under the null hypothesis.
Statement of the problem

Given a random sample \( \{X_1, \ldots, X_n\} \) of \( X \) with cdf \( F \), the goal is to test:

\[
H_0 : \ F \in \mathcal{F} = \{F_\theta\}_{\theta \in \Theta \subset \mathbb{R}^q}, \quad \text{vs.} \quad H_a : \ F \not\in \mathcal{F}
\]
Kolmogorov-Smirnov and Cramer-Von Mises tests

\[ T_n(F_n, F_{\hat{\theta}}) = \begin{cases} 
\sup_x |F_n(x) - F_{\hat{\theta}}(x)| \\
\int (F_n(x) - F_{\hat{\theta}}(x))^2 dF_n(x)
\end{cases} \]

where \( F_n(x) = \frac{\#\{j | X_{ij} \leq x\}}{n} \) is the empirical distribution and \( F_{\hat{\theta}} \) is a parametric estimation, being \( \hat{\theta} \) a \( \sqrt{n} \)-consistent estimator of \( \theta \in \Theta \subset \mathbb{R}^q \).
\( \chi^2 \) test

\[
T_n(\hat{f}_H, f_{\theta}) = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \approx n \left( \int \frac{\hat{f}_H^2(x)}{f_{\theta}(x)} dx - 1 \right)
\]

with \( \mathcal{X} = \text{Support}(F) = \bigcup_{i=1}^{k} A_i \) and

\[
O_i = \#\{j | X_j \in A_i\}, \quad E_i = n \int_{A_i} f_{\theta}(x) dx \quad i = 1, \ldots, k
\]

\[
\hat{f}_H(x) = \sum_{i=1}^{k} \frac{O_i}{n \, l(A_i)} \mathbb{I}(x \in A_i)
\]

where \( l(\cdot) \) denotes the Lebesgue measure of \( A_i \) and \( \mathbb{I}(\cdot) \) is the indicator function.
Two basic references

Durbin (1973) and Bickel and Rosenblatt (1973) settled the beginnings of the mathematical developments for GOF tests, based on the estimation of the cdf and the density function, respectively.
Durbin (1973)

\[ T_n = T(\alpha_n) = T \left( \sqrt{n}(F_n(\cdot) - F_{\hat{\theta}}(\cdot)) \right). \]

Hence, \( \alpha_n = \sqrt{n}(F_n(\cdot) - F_{\hat{\theta}}(\cdot)) \) is the empirical process with parametric estimation for the cdf.

Some examples:

\[ T_n = \sup_x |\alpha_n(x)| \]

\[ T_n = \int \alpha_n^2(x) dF_n(x) \]

\( T_n \) any continuous functional of \( \alpha_n \)

Note that...

The limit convergence of \( \alpha_n \) is given by a Gaussian process where the covariance structure of \( \hat{\theta} \) is directly involved and tabulation is needed.
Bickel and Rosenblatt (1973)

\[ T_n = T(\tilde{\alpha}_n) = T \left( \sqrt{n h} (f_{nh}(\cdot) - \mathbb{E}_{\hat{\theta}} f_{nh}(\cdot)) \right), \]

where \( f_{nh}(x) = (nh)^{-1} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right) \) is the kernel density estimator, with \( K \) the kernel function and \( h \) smoothing parameter (c.f. Parzen (1962) and Rosenblatt (1956)).

Note that...

The limit convergence of \( \tilde{\alpha}_n \) is given by a Gaussian process where the covariance structure of \( \hat{\theta} \) is NOT involved (c.f. Rosenblatt (1990)).
For instance, consider the $p$-dimensional case with test statistic:

$$T_n = \int \tilde{\alpha}_n^2(x)\omega(x)dx = \int \left[\sqrt{nh^p}(f_{nh}(x) - \mathbb{E}_{\hat{\theta}}(f_{nh}(x)))\right]^2 \omega(x)dx.$$

It holds that:

$$h^{-p/2}\left(T_n - \int K^2 \cdot \int f \omega\right) \xrightarrow{d} N\left(0, 2\int (K * K)^2 \int f^2 \omega^2\right)$$

as $h \equiv h_n \to 0$ and $nh^p_n \to \infty$ (c.f. Fan (1994, 1998)).
Introduction
Density and regression

- Durbin (1973)
- Bickel and Rosenblatt (1973)
- Ahmad and Cerrito (1993)
- Fan (1994, 1998)
- Gouriéroux and Tenreiro (2001)
- Neumann and Paparoditis (2000)
- Lee and Na (2002)
- Giné and Mason (2004)
- Chebana (2004)
- Cao and Lugosi (2005)
- Bachmann and Dette (2005)
- Chebana (2006)
- Liang and King (2007)
- Tenreiro (2007, 2009) ... etc.
Extension

These ideas have been extended in the nineties to the general case of a regression model (c.f. Härdle and Mammen (1993), GM and Cao (1993)).

Given a regression model (fixed or random design):

\[ Y_i = m(X_i) + \epsilon_i, \quad i = 1, \ldots, n \]

where \( \mathbb{E}(Y_i | X_i) = m(X_i) \). The goal is to test

\[ H_0 : m \in \mathcal{M} = \{m_\theta \} \theta \in \Theta \subseteq \mathbb{R}^q, \quad \text{vs.} \quad H_a : m \notin \mathcal{M} \]

with \( m(x) = \mathbb{E}(Y | X = x) \) the regression function of \( Y \) over \( X \),
\( \sigma^2 = \text{Var}(Y | X = x) \) and \( f \) the density of the explanatory variable (if exists).
Introduction

Density and regression

$X$: explanatory variable
$Y$: response random variable

Regression function: $m(x) = \mathbb{E}(Y \mid X = x)$

Question...

Is the model $m \in \{m_\theta; \theta \in \Theta\}$ enough well supported by the data?

See Seber (1977) for the linear regression case ($m_\theta(\cdot) = A^t(\cdot)\theta$) and Seber and Wild (1989) for nonlinear $m_\theta$. 
Polynomial regression

\( X \) is a one dimensional random variable.

\[ A^t(x) = (1, x, x^2, \ldots, x^{q-1}) \in \mathbb{R}^q \]

\[ \theta = (\theta_1, \theta_2, \ldots, \theta_q)^t \in \Theta \subset \mathbb{R}^q \]

Multiple regression

\( X \) is a \( q \)-dimensional random variable.

\[ A^t(x) = x \in \mathbb{R}^q, \quad \theta \in \Theta \subset \mathbb{R}^q \]
Nonlinear regression

\[ X = (X_1, X_2)^t. \]

\[ \theta = (\theta_1, \theta_2)^t \in \Theta = \{(s, t) \in \mathbb{R}^2 / s + t \neq 0\} \]

\[ m_{\theta}(x) = \frac{1}{\theta_1 + \theta_2} \exp(\theta_1 x_1 + \theta_2 x_2) \]
A simulated example:

- Model:
  \[ Y = 2x^2 - 5x + \cos(2\pi x) + \varepsilon \]
- \( n = 500 \)
- \( \varepsilon \sim \mathcal{N}(0, 1) \)
- Null hypothesis:
  \[ H_0 : m \text{ is linear} \]
A real dataset:

- 235 observations
- Early eighties
- $X$: income
- $Y$: expenditure on food for Belgian working class households.
- Null hypothesis:

\[ H_0 : m \text{ is linear} \]
Some other features to take into account:

- Fixed or random design

- Nonparametric estimator $m_{nh}(x) = \sum_{i=1}^{n} W_{ni}(x)Y_i$:
  - Nadaraya-Watson
  - Local Polynomial
  - Priestley-Chao
  - Splines
  - Orthogonal expansions
  - Wavelets
  - . . .
The local polynomial regression (c.f. Fan and Gijbels (1996))

\[ m_{nh}(x) = \hat{\beta}_0(x) = \sum_{j=1}^{n} W_{n,\bar{q}} \left( \frac{x - X_i}{h} \right) Y_i, \]

where \( \hat{\beta}(x) = (\hat{\beta}_0(x), \ldots, \hat{\beta}_{\bar{q}}(x))^t \), is the minimizer of:

\[
\sum_{i=1}^{n} \left( Y_i - \sum_{r=0}^{\bar{q}} \beta_r (x - X_i)^r \right)^2 K \left( \frac{x - X_i}{h} \right),
\]

and \( W_{n,\bar{q}} = u^t (XX^T)^{-1} (1, ht, \ldots, h^{\bar{q}+t\bar{q}}) \frac{K(t)}{h} \), with \( u^t = (1, 0, \ldots, 0) \in \mathbb{R}^{\bar{q}+1}, X = ((x - X_i)^j)_{1 \leq i \leq n, 1 \leq j \leq \bar{q}}, W = \text{diag} \left( K \left( \frac{x-X_i}{h} \right) \right) \).
The initial empirical process, for the $p$-dimensional case in the explanatory variable, is given by:

$$\overline{\alpha}_n(x) = \sqrt{n h^p} \left( m_{nh}(x) - \mathbb{E}_{\hat{\theta}}(m_{nh}(x)) \right)$$

$$= \sqrt{n h^p} \sum_{i=1}^{n} W_{ni}(x) (Y_i - m_{\hat{\theta}}(X_i))$$

$$= \sqrt{n h^p} \sum_{i=1}^{n} W_{ni}(x) \hat{\varepsilon}_i$$

where $\mathbb{E}_{\hat{\theta}}$ is the estimation of $\mathbb{E}_{\theta_0}$ (with $\theta_0$ theoretical parameter under $H_0$) and $\hat{\theta}$ is a $\sqrt{n}$-consistent estimator of $\theta_0$ (for instance, least squares, maximum likelihood, . . .).
For instance, in order to test a polynomial regression model:

\[ H_0 : m(x) = \sum_{j=1}^{q} \theta_j x^{j-1}, \quad \text{vs.} \quad H_a : m(x) \neq \sum_{j=1}^{q} \theta_j x^{j-1} \]

we may consider \( T_n = \int \alpha_n^2(x) \omega(x) \, dx \). It holds that:

\[ h^{-1/2} (T_n - c_1) \xrightarrow{d} N(0, c_2) \]

where

\[ c_1 = \int \tilde{K}^2(x) \, dx \int \frac{\sigma^2(x) \omega(x)}{f(x)} \, dx, \quad c_2 = 2 \int (\tilde{K} \ast \tilde{K})^2(x) \, dx \int \frac{\sigma^4(x) \omega^2(x)}{f^2(x)} \, dx \]

(Alcalá, Cristóbal and GM, 1999)
Drawbacks of the smoothing approach

- Bandwidth choice.
- Slow rate of convergence for $T_n$ to its Gaussian limit.
- Unknown curves involved in the test statistic must be estimated.
The previous technique for testing regression models is based on the ideas of the smoothing methods from Bickel and Rosenblatt (1973) for the density function.

**The integrated regression function**

In a similar way to the tests for the distribution function $F(x) = \int_{-\infty}^{x} f(t) dt$, we may consider the integrated regression function:

$$I(x) = \int_{-\infty}^{x} m(t)dF(t) = \mathbb{E}(Y \cdot \mathbb{I}(X \leq x))$$
The integrated regression function can be nonparametrically estimated as follows:

\[ I_n(x) = \frac{1}{n} \sum_{i=1}^{n} Y_i \cdot \mathbb{I}(X_i \leq x) \]

with associated empirical process:

\[ \overline{\alpha}_n(x) = \sqrt{n}(I_n(x) - \mathbb{E}_{\hat{\theta}}(I_n(x))) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \mathbb{I}(X_i \leq x)\hat{\epsilon}_i \]

This empirical process will be the basis for a broad class of test statistics (Stute, 1997).
In the general testing problem:

\[ H_0 : g \in \mathcal{G} = \{ g_\theta \}_{\theta \in \Theta}, \quad \text{vs.} \quad H_a : g \notin \mathcal{G} = \{ g_\theta \}_{\theta \in \Theta} \]

with test statistic \( T_n = T(g_n, g_\hat{\theta}) \) where \( g \) may be \( F_\theta, f_\theta, m_\theta \) or \( I_\theta \), calibration of critical points is crucial.

**Critical point calibration**

Estimate \( c_\alpha \) such that:

\[ \mathbb{P}_{H_0}(T_n \geq c_\alpha) = \alpha \]
How can we estimate $c_\alpha$?

- Using the asymptotic normality (Case $g_n = f_{nh}$ or $g_n = m_{nh}$).
- Approximating the distribution of the empirical process $\alpha_n$ or $\overline{\alpha_n}$ (Case $g_n = F_n$ or $g_n = I_n$).
- Using Bootstrap.
- ...
In Stute et al. (1993), the critical point $c_\alpha$ is estimated using parametric Bootstrap.

Specifically, if $T_n^* = T(F_n^*, F_{\hat{\theta}}^*)$, $c_\alpha$ is estimated by $\hat{c}_\alpha$ such that:

$$\mathbb{P}^*_{H_0}(T_n^* \geq \hat{c}_\alpha) = \alpha$$

where $\mathbb{P}^*$ is the probability under resampling and $F_n^*$ and $F_{\hat{\theta}}^*$ are obtained with the Bootstrap samples $\{X_1^*, \ldots, X_n^*\}$ (iid from $X^* \sim F_{\hat{\theta}}$).
Some examples

- \( T_n = \int \frac{\alpha_n^2(x)}{\alpha_n^2(x)\omega(x)} dx \), in regression models with kernel estimator (Härdle and Mammen, 1993).
- \( T_n = \int \frac{\alpha_n^2(x)}{\alpha_n^2(x)\omega(x)} dx \), with empirical regression processes (Stute et al., 1998).
- Naive Bootstrap
- Wild Bootstrap
- . . .
The bootstrap resamples \( \{(X_i^*, Y_i^*)\}_{i=1}^n \) are obtained as:

1. **Construct the parametric residuals:**
   \[
   \hat{\varepsilon}_i = Y_i - m_{\hat{\theta}}(X_i), \quad i = 1, 2, \ldots, n.
   \]

2. **Recenter the previous residuals:**
   \[
   \bar{\hat{\varepsilon}}_i = \hat{\varepsilon}_i - \bar{\hat{\varepsilon}}, \quad i = 1, 2, \ldots, n, \text{ where } \bar{\hat{\varepsilon}} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i}{n}.
   \]

3. **Draw bootstrap versions of the residuals, \( \varepsilon_i^* \), from the empirical cdf of the \( \{\bar{\hat{\varepsilon}}_i\}_{i=1}^n \).**

4. **Compute**
   \[
   Y_i^* = m_{\hat{\theta}}(X_i) + \varepsilon_i^*, \quad i = 1, 2, \ldots, n \text{ (no resampling of the } X\text{'s).}
   \]

1. Construct the parametric residuals:

\[ \hat{\varepsilon}_i = Y_i - m_\hat{\theta}(X_i), \ i = 1, 2, \ldots, n. \]

2. Draw independent r.v. \( V_1^*, V_2^*, \ldots, V_n^* \) (also independent of the observed sample) satisfying

\[ E^*(V_i^*) = 0, \ E^*(V_i^{*2}) = 1, \ E^*(V_i^{*3}) = 1 \]

and construct the \( \varepsilon_i^* = \hat{\varepsilon}_i V_i^*. \)

3. Compute \( Y_i^* = m_\hat{\theta}(X_i) + \varepsilon_i^*, \ i = 1, 2, \ldots, n \) (no resampling of the \( X \)'s).
Other alternative techniques

- Monte Carlo (Zhu, 2005)
Pitman alternatives

\[ H_0 : \ m(\cdot) = m_\theta(\cdot), \ \theta \in \Theta, \ \ vs. \ H_{1n} : \ m_n(\cdot) = m_\theta(\cdot) + c_n d(\cdot) \]

where \( c_n \to 0 \) and \( d \) denotes the direction of the alternative.

A test with the power higher than \( \alpha \) and with the fastest \( c_n \) tending to zero will be the most powerful.
Summary of the asymptotic sensitivity ($c_n$) to deviations from the hypothesis for Pitman alternatives:

<table>
<thead>
<tr>
<th>Test</th>
<th>$H_0$</th>
<th>$H_1$</th>
<th>$c_n$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F$-test</td>
<td>Param.</td>
<td>Param.</td>
<td>$n^{-1/2}$</td>
</tr>
<tr>
<td>Raw residuals</td>
<td>Param.</td>
<td>Nonparam.</td>
<td>$n^{-1/4}$</td>
</tr>
<tr>
<td>Smoothing residuals</td>
<td>Param.</td>
<td>Nonparam.</td>
<td>$n^{-1/2}h^{-1/4}$</td>
</tr>
<tr>
<td>Empirical process</td>
<td>Param.</td>
<td>Nonparam.</td>
<td>$n^{-1/2}$</td>
</tr>
</tbody>
</table>

In the general case of a $p$-dimensional model, the rate would be (based on residuals) $c_n = n^{-1/2}h^{-p/4}$. 
Based on the previous slide, we may conclude that test based on regression empirical process are better in practice. FALSE!

Some simulation studies (c.f. Miles and Mora, 2002) show that both families may be competitive in finite samples. In addition, the modified test statistic:

\[
T_n = \max_{h \in H_n} \frac{\int \alpha_n h^2(x) \omega(x) dx - \hat{E}_{H_0} \left( \int \alpha_n h^2(x) \omega(x) dx \right)}{\hat{\text{Var}}_{H_0} \left( \int \alpha_n h^2(x) \omega(x) dx \right)}
\]

where $H_n$ is a family of smoothing parameters, can detect alternatives at rate $c_n \sim n^{-1/2} (\log(\log n))^{1/2}$ (Horowitz and Spokoiny, 2001).

In a regression model, assume that $\varepsilon \sim N(0, \sigma^2)$. The GLRT statistic is given by:

$$T_n = l(m_{nh}, \hat{\sigma}) - l(m_\hat{\theta}, \hat{\sigma}_0)$$

where $l(m, \sigma)$ is the Gaussian log-likelihood:

$$l(m, \sigma) = -n \log(\sqrt{2\pi\sigma^2}) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (Y_i - m(X_i))^2$$

and

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - m_{nh}(X_i))^2, \quad \hat{\sigma}_0^2 = \frac{1}{n} \sum_{i=1}^{n} (Y_i - m_\hat{\theta}(X_i))^2$$
Chen *et al.*, 2003, Chen and Van Keilegom, 2009

The test statistic is:

\[ T_n = \int \left( -2 \log(L_n(\tilde{m}(x, \hat{\theta}))n^n) \right) \omega(x) dx \]

where

\[ L_n(\tilde{m}(x, \hat{\theta})) = \max \prod_{i=1}^{n} p_i(x) \]

subject to \( \sum_{i=1}^{n} p_i(x) = 1 \) and

\[ \sum_{i=1}^{n} p_i(x) K \left( \frac{x - X_i}{n} \right) (Y_i - \tilde{m}(x, \hat{\theta})) = 0, \quad \text{with} \quad \tilde{m}(x, \hat{\theta}) = \mathbb{E}_{\hat{\theta}}(m_{nh}(x)) \]
For $H_0: m \in \mathcal{M} = \{m_\theta\}_{\theta \in \Theta \subset \mathbb{R}^p}$, this test verifies that:

$$h^{-p/2}(T_n - c_1) \xrightarrow{d} N(0, c_2),$$

where $c_1 = 1$ and

$$c_2 = 2 \frac{\int (K * K)^2(x) dx}{\int K^2(x) dx} \frac{\int \omega(x) dx}{\int \omega(x) dx}.$$


Van Keilegom, Sánchez-Sellero and GM, 2008

Take into account that:

$$H_0 : m \in \{m_\theta\}_{\theta \in \Theta} \iff \mathbb{E}(\mathbb{I}(X \leq x)(Y - m_{\theta_0}(X))) = 0,$$

for $\theta_0 \in \Theta$ and $x \in \text{Support}(X)$, and the empirical likelihood for $p = 1$:

$$L(F) = \prod_{i=1}^{n} \left(F(X_i, Y_i) - F(X_i^-, Y_i) - F(X_i, Y_i^-) + F(X_i^-, Y_i^-)\right)$$

with $\{(X_i, Y_i)\}_{i=1}^{n}$ iid from $(X, Y)$ with joint distribution $F$
The test statistic based on the empirical likelihood ratio is:

\[
\Lambda_n(x) = \sup \left\{ \frac{L(F)}{\sup L(F)} : \mathbb{E}_F \left( \mathbb{I}(X \leq x) (Y - m_{\hat{\theta}}(X)) \right) = 0 \right\}
\]

\[
= \sup \left\{ n^n \prod_{i=1}^{n} p_i ; \ p_i \geq 0, \ i = 1, \ldots, n, \ \sum_{i=1}^{n} p_i = 1, \right. \\
\left. \sum_{i=1}^{n} p_i \mathbb{I}(X_i \leq x)(Y_i - m_{\hat{\theta}}(X_i)) = 0 \right\}
\]

where \( T_n \) may be any continuous functional of \( \Lambda_n(\cdot) \).
In a location-scale nonparametric framework, a regression model can be expressed as:

\[ Y = m(X) + \sigma(X) \varepsilon, \]

- \( m(x) = \mathbb{E}(Y|X = x) \) smooth regression function
- \( \sigma^2(x) = \text{Var}(Y|X = x) \) variance function
- \( \varepsilon \) error (independent of the covariate) with distribution function:

\[ F_{\varepsilon}(y) = \mathbb{P}(\varepsilon \leq y) = \mathbb{P}\left( \frac{Y - m(X)}{\sigma(X)} \leq y \right) \]

- Sample: \((X_i, Y_i), i = 1, \ldots, n\) iid observations from \((X, Y)\)
Akritas and Van Keilegom (2001) proposed estimating the error distribution by the empirical distribution of estimated residuals:

\[
\hat{F}_\varepsilon(y) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I} \left( \frac{Y_i - \hat{m}(X_i)}{\hat{\sigma}(X_i)} \leq y \right),
\]

where \( \hat{m}(x) \) and \( \hat{\sigma}^2(x) \) are Nadaraya-Watson estimators of \( m \) and \( \sigma^2 \).

\[
\hat{m}(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i, \quad \text{and} \quad \hat{\sigma}^2(x) = \sum_{i=1}^{n} W_{ni}(x) Y_i^2 - \hat{m}^2(x),
\]

Nadaraya-Watson weights: \( W_{ni}(x) = \frac{K_h((x - X_i)/h)}{\sum_{j=1}^{n} K_h((x - X_j)/h)} \)
During the last few years, the estimation of the error distribution has been used to test several hypothesis about the regression model. The basic idea of the tests is the following:

1. Estimate the error distribution nonparametrically: \( \hat{F}_\varepsilon \).
2. Estimate the error distribution introducing the null hypothesis: \( \hat{F}_{\varepsilon 0} \).
3. Compare the two estimates of the error distribution.
4. Use Bootstrap to approximate the critical values of the test.
An important example is the GOF test for parametric regression models:

\[ H_0 : \ m \in \mathcal{M} = \{m_\theta\}_{\theta \in \Theta}, \text{ vs. } H_a : \ m \notin \mathcal{M} = \{m_\theta\}_{\theta \in \Theta} \]

- Van Keilegom, GM and Sánchez-Sellero, 2008 (for \( p = 1 \))
- Neumeyer, N. and Van Keilegom, 2010 (for \( p \geq 1 \))
The equality of the error distribution can be also interpreted in terms of the characteristic functions. For instance, based on functionals of the empirical characteristic functions, the following test statistic can be designed:

\[ T_n = n \int |\hat{\phi}_\varepsilon(t) - \hat{\phi}_{\varepsilon_0}(t)|^2 \omega(t) dt \]

where \( \hat{\phi}_\varepsilon(t) = n^{-1} \sum_{j=1}^{n} e^{it\hat{\varepsilon}_j} \) and \( \hat{\phi}_{\varepsilon_0}(t) = n^{-1} \sum_{j=1}^{n} e^{it\hat{\varepsilon}_0} \).

- Jiménez–Gamero et al. (2005)
The curse of dimensionality

Both in tests based on $\overline{\alpha_n}$ or on $\overline{\alpha_n}$, the curse of dimensionality as $p$ increases (being $p$ the dimension of the explanatory variable) can be appreciated.

- For tests based on $\overline{\alpha_n}$, the effect of the increasing dimension is clear when regarding the asymptotic power.
Inspired on the projection pursuit ideas, the null hypothesis $H_0 : m \in \mathcal{M}_\theta$ is true if and only if $m = m_{\theta_0} \in \mathcal{M}_\theta$, and this is also equivalent to $E(\varepsilon | X) = E(\varepsilon_0 | X) = E(Y - m_{\theta_0}(X) | X) = 0$. In addition, this is also equivalent to:

$$\sup_{\beta, \|\beta\|=1} \sup_{\nu} |E(\varepsilon | \beta^t X = \nu)| = 0 \Leftrightarrow \sup_{\beta, \|\beta\|=1} E(\varepsilon E(\varepsilon | \beta^t X)) = 0$$

under some regularity conditions, and this allows for the construction of:

$$T_n = \sup_{\beta, \|\beta\|=1} \sum_{i<j} K_h(\beta^t (X_i - X_j))(Y_i - m_{\hat{\theta}}(X_i))(Y_j - m_{\hat{\theta}}(X_j))$$

➤ Lavergne and Patilea (2008)
Another interesting idea consists in projecting the covariate $X$ in the direction of $\beta = \beta_0$ such that this $\beta_0$ (with $\|\beta_0\| = 1$) minimizes $\mathbb{E}^2(\varepsilon - \mathbb{E}(\varepsilon|\beta^t X)) = \mathbb{E}^2(\varepsilon - m_\beta(X))$. This enables to construct test statistics such as

$$T_n = \frac{1}{n} \sum_{i=1}^{n} \omega(X_j) \left( \hat{\varepsilon}_{j0} - \hat{m}_{\hat{\beta}_j}(\hat{\beta}_j^t X_j) \right)^2$$

where

$$\hat{\beta}_j = \arg \min_{\beta, \|\beta\|=1} \sum_{i \neq j} \left( \hat{\varepsilon}_{i0} - \hat{m}_\beta^j(X_i) \right)^2, \quad j = 1, \ldots, n$$

being

$$\hat{m}_\beta^j(x) = \frac{1}{n \hat{f}_\beta^j(X_j)} \sum_{i \neq j} K_h(\beta^t(X-x_i)) \hat{\varepsilon}_{i0}, \quad \text{and} \quad \hat{f}_\beta^j(x) = \frac{1}{n} \sum_{i \neq j} K_h(\beta^t(x-X_i))$$

Xia (2009)
Regarding the tests based on empirical regression processes, the empirical process $\overline{\alpha_n}$ can be replaced by

$$\overline{\alpha_n}^g(t) = n^{-1/2} \sum_{i=1}^{n} (g(X_i) - \bar{g})I(\hat{\varepsilon}_i \leq t), \quad t \in \mathbb{R}$$

indexed unidimensionally in $t$, with $\bar{g} = n^{-1} \sum_{i=1}^{n} g(X_i)$.

➤ Stute et al. (2008)
The almost sure characterization of the null hypothesis as $\mathbb{E}(\varepsilon_0 I(\beta^t X \leq u)) = 0$, for some $\theta_0 \in \Theta$, $\forall u \in \mathbb{R}$ and $\forall \beta$ such that $\|\beta\| = 1$ leads to a process $\overline{\alpha}_n(\beta, u) = n^{-1/2} \sum_{i=1}^{n} \hat{\varepsilon}_i I(\beta^t X_i \leq u)$, indexed in $\beta$ and $u$.

$$T_{nKS} = \sup_u \sup_{\beta, \|\beta\|=1} \left| n^{-1/2} \sum_{i=1}^{n} \hat{\varepsilon}_i I(\beta^t X_i \leq u) \right| = \sup_u \sup_{\beta, \|\beta\|=1} |\overline{\alpha}_n(\beta, u)|$$

$$T_{nCM} = \int_{\mathbb{S}^p \times \mathbb{R}} (\overline{\alpha}_n(\beta, u))^2 dF_{n\beta}(u) d\omega(\beta)$$

being $F_{n\beta}$ the empirical distribution of $\{\beta^t X_i\}_{i=1}^{n}$ and $\omega$ a weight function over the projection direction.

All the previous GOF tests can be extended in several ways:
- Extending the null hypothesis
- Incomplete data
- Dependent data
We may consider the following null hypothesis:

- Partial linear models:

\[ H_0 : Y_i = X_i^t \theta + m(Z_i) + \varepsilon_i, \ i = 1, \ldots, n \]

- Generalized partial linear models:

\[ H_0 : \mathbb{E}(Y_i | X_i, Z_i) = G(X_i^t \theta + m(Z_i)) \]

with \( G \) a known link function.

- Significance test:

\[ H_0 : \mathbb{E}(Y_i | X_i, Z_i) = \mathbb{E}(Y_i | X_i) \]

- Testing additivity:

\[ \mathbb{E}(Y_i | X_{i1}, \ldots, X_{ip}) = \sum_{j=1}^{p} m_j(X_{ij}) \]
GOF tests for regression models: the functional data case

- GOF in other contexts and some recent advances
- Extending the null hypothesis

- Goodness–of–Fit Tests for Interest Rate Models
- Goodness–of–Fit Tests for Directional Data
- Goodness–of–Fit Tests for Functional Data
Part I: Review of Goodness–of–Fit tests for regression models
1. Introduction
2. Calibration, size and power
3. Recent GOF tests for regression models
4. GOF in other contexts and some recent advances

Part II: The functional case: GOF of the functional linear model
1. Functional data
2. The test
3. Simulation study
4. Real data application
Functional data

Each observation is a curve! Real examples in:
- Meteorology: temperature, precipitation, wind speed, . . .
- Finance: evolution of asset prices/returns.
- Spectrometry
- . . .
Figure: Tecator dataset. Absorbance coloured by fat content.
Figure: Yearly evolution of temperatures in 76 AEMET monitoring stations.
Figure: SP500 index. Each day is a functional datum.
Objective

Propose a Goodness–of–Fit test for the null hypothesis of the functional linear model,

\[ Y = \langle \mathcal{X}, \beta \rangle + \varepsilon = \int \mathcal{X}(t)\beta(t)dt + \varepsilon, \]

with \( \varepsilon \) a centred r.v. independent from \( \mathcal{X} \). This is equivalent to test \( H_0 : m(\cdot) \in \{ \langle \cdot, \beta \rangle : \beta \in \mathcal{H} \} \), being \( \mathcal{H} = \mathcal{L}_2[0, T] \) the Hilbert space of square integrable functions.

Let $\mathcal{X}(t)$ be a functional r.v. taking values in $\mathbb{H} = \mathcal{L}_2[0, T]$. Let $\{\Psi_j\}_{j=1}^{\infty}$ be a basis of $\mathbb{H}$. Then for each observation $\mathcal{X}_i$, $i = 1, \ldots, n$ of $\mathcal{X}$, we can express

$$\mathcal{X}_i = \sum_{j=1}^{\infty} x_{ij} \Psi_j.$$ 

Let $\{\Psi_j\}_{j=1}^{p}$ the $p$–truncate basis with the first $p$ elements of $\{\Psi_j\}_{j=1}^{\infty}$. The representation of $\mathcal{X} \in \mathbb{H}$ in this truncated basis is denoted by

$$\mathcal{X}_i^{(p)} = \sum_{j=1}^{p} x_{ij} \Psi_j.$$ 

The basis choice and the number of elements of the truncated basis are crucial in order to capture correctly the information given by the functional process.
The Functional Linear Model states that
\[
Y = \langle \mathcal{X}, \beta \rangle + \varepsilon = \int \mathcal{X}(t)\beta(t)dt + \varepsilon,
\]
with $\varepsilon$ a centred r.v. independent from $\mathcal{X}$.

The estimation of the functional parameter $\beta$ is done by minimising the RSS:
\[
\hat{\beta} = \arg\min_{\beta \in \mathcal{H}} \sum_{i=1}^{n} (Y_i - \langle \mathcal{X}_i, \beta \rangle)^2.
\]

Different methods have been proposed to search for the $\hat{\beta}$ that minimizes the RSS (see Ferraty and Romain (2011)):

1. Using a basis representation of B–splines or Fourier functions.
2. Using Principal Components (PC).
Let $(\mathcal{X}, Y)$ be r.v.’s in $\mathbb{H} \times \mathbb{R}$ and the sample $\{(\mathcal{X}_i, Y_i)\}_{i=1}^n$.

We want to test this null composite hypothesis:

$$H_0 : m \in \{\langle \cdot, \beta \rangle : \beta \in \mathbb{H}\},$$

versus the general alternative

$$H_1 : \mathbb{P}\{m \notin \{\langle \cdot, \beta \rangle : \beta \in \mathbb{H}\}\} > 0.$$

The simple hypothesis, i.e. checking for a specific functional linear model, is also of interest

$$H_0 : m (\mathcal{X}) = \langle \mathcal{X}, \beta_0 \rangle,$$

for a fixed $\beta_0 \in \mathbb{H}$

and it includes the important case of no interaction between the functional covariate and the scalar response ($\beta_0 = 0$).
Lemma

Let $\beta \in \mathbb{H}$. The following statements are equivalent:

I  $m(\mathcal{X}) = \langle \mathcal{X}, \beta \rangle$, $\forall \mathcal{X} \in \mathbb{H}$.

II  $\mathbb{E} [Y - \langle \mathcal{X}, \beta \rangle | \mathcal{X} = x] = 0$, for a.e. $x \in \mathbb{H}$.

III  $\mathbb{E} [Y - \langle \mathcal{X}, \beta \rangle | \langle \mathcal{X}, \gamma \rangle = u] = 0$, for a.e. $u \in \mathbb{R}$ and $\forall \gamma \in S_{H}$.

III'  $\mathbb{E} [Y - \langle \mathcal{X}, \beta \rangle | \langle \mathcal{X}, \gamma \rangle = u] = 0$, for a.e. $u \in \mathbb{R}$ and $\forall \gamma \in S^{p}_{H}$, $\forall p \geq 1$.

IV  $\mathbb{E} [(Y - \langle \mathcal{X}, \beta \rangle ) 1_{\{\langle \mathcal{X}, \gamma \rangle \leq u\}}] = 0$, for a.e. $u \in \mathbb{R}$ and $\forall \gamma \in S_{H}$.

IV'  $\mathbb{E} [(Y - \langle \mathcal{X}, \beta \rangle ) 1_{\{\langle \mathcal{X}, \gamma \rangle \leq u\}}] = 0$, for a.e. $u \in \mathbb{R}$ and $\forall \gamma \in S^{p}_{H}$, $\forall p \geq 1$.

Lemma based on the results of Patilea et al. (2012).
A possible way to measure the deviation of the data from \( H_0 \) is by the projected empirical process:

\[
R_n(u, \gamma) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left( Y_i - \langle X_i, \hat{\beta} \rangle \right) \mathbb{1}\{\langle x_i, \gamma \rangle \leq u\}.
\]

To measure the distance of the empirical process from zero: Cramér–von Mises and Kolmogorov–Smirnov norms, adapted to the projected space \( \Pi = \mathbb{R} \times S_{\mathbb{H}} \):

\[
PC_{\text{vM}}_n = \int_{\Pi} R_n(u, \gamma)^2 F_{n, \gamma}(du) \omega(d\gamma),
\]

\[
PK_{\text{S}}_n = \sup_{(u, \gamma) \in \Pi} |R_n(u, \gamma)|,
\]

where \( F_{n, \gamma} \) is the ecdf of \( \{\langle X_i, \gamma \rangle\}_{i=1}^{n} \) and \( \omega \) represents a functional measure on \( S_{\mathbb{H}} \).
A $p$–truncated version of this statistic is:

$$R_{n,p} \left( u, \gamma^{(p)} \right) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left( Y_i - x_{i,p}^T \Psi b_p \right) \mathbb{1}\{x_{i,p}^T \Psi g_p \leq u\} = R_{n,p} \left( u, g_p \right),$$

where $b_p$ are the coefficients of $\hat{\beta}$ in the $p$–truncated basis $\{\Psi_j\}_{j=1}^p$, $\Psi = (\langle \Psi_i, \Psi_j \rangle)_{i,j}$ ($I_p$ if the basis is orthonormal) and $g_p$ are the coefficients of $\gamma$ in the truncated basis.

A simplified version of the statistics PCvM$_n$ considering the uniform distribution of the sphere is:

$$\text{PCvM}_{n,p} = \int_{S^p \times \mathbb{R}} |R|^{-1} R_{n,p}(u, R^{-1} g_p)^2 F_{n,R^{-1} g_p}(du) dg_p,$$

where $R$ is the $p \times p$ matrix such that $\Psi = R^T R$. 
By calculus analogous to those of Escanciano (2006),

\[ PCvM_{n,p} = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{r=1}^{n} \varepsilon_i \varepsilon_j A_{ijr}, \]

where:

\[ A_{ijr} = A^{(0)}_{ijr} \frac{\pi^{p/2-1}}{\Gamma \left( \frac{p}{2} + 1 \right)} |R|^{-1}, \]

\[ A^{(0)}_{ijr} = \begin{cases} 
2\pi, & x'_{i,p} = x'_{j,p} = x'_{r,p}, \\
\pi, & x'_{i,p} = x'_{j,p}, x'_{i,p} = x'_{r,p} \text{ or } x'_{j,p} = x'_{r,p}, \\
\pi - \arccos \left( \frac{(x'_{i,p}-x'_{r,p})^T (x'_{j,p}-x'_{r,p})}{\|x'_{i,p}-x'_{r,p}\| \cdot \|x'_{j,p}-x'_{r,p}\|} \right), & \text{else.} 
\end{cases} \]
Test procedure for simple hypothesis

Let be \( \{(X_i, Y_i)\}_{i=1}^n \) an iid sample:

1. Express \( X_i(\cdot) \) (and optionally \( \beta_0(\cdot) \)) in a \( p \)-truncated basis.
2. Construct \( \epsilon_i = Y_i - \langle X_i, \beta_0 \rangle \).
3. Compute \( \text{PCvM}_{n,p} = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \sum_{r=1}^n \epsilon_i \epsilon_j A_{ijr} \).
4. Bootstrap procedure (Wild bootstrap):
   - Construct \( \epsilon_i^* = V_i \epsilon_i \), with \( \mathbb{E} [V_i] = 0, \mathbb{E} [V_i^2] = 1 \).
   - Compute \( \text{PCvM}_{n,p}^* = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \sum_{r=1}^n \epsilon_i^* \epsilon_j^* A_{ijr} \).
5. \( p \)-value \( \approx \# \left\{ \text{PCvM}_{n,p} \leq \text{PCvM}_{n,p}^* \right\} / B \).
Test procedure for composite hypothesis

Let be \( \{(\mathcal{X}_i, Y_i)\}_{i=1}^n \) an iid sample:

1. Set \( \mathcal{X}_i(\cdot) := \mathcal{X}_i(\cdot) - \bar{\mathcal{X}}_i(\cdot) \) and \( Y_i := Y_i - \bar{Y} \).
2. Express \( \mathcal{X}_i(\cdot) \) and \( \beta(\cdot) \) in a \( p \)-truncated basis (bspline, PC or PLS).
3. Construct \( \epsilon_i = Y_i - \langle \mathcal{X}_i, \hat{\beta} \rangle \).
4. Compute \( \text{PCvM}_{n,p} = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \sum_{r=1}^n \epsilon_i \epsilon_j A_{ijr} \).
5. Bootstrap procedure (Wild bootstrap):
   - Construct \( Y_i^* = \langle \mathcal{X}_i, \hat{\beta} \rangle + \epsilon_i^*, \) where \( \epsilon_i^* = V_i \epsilon_i \), with \( \mathbb{E}[V_i] = 0 \), \( \mathbb{E}[V_i^2] = 1 \).
   - Estimate \( \beta^*(\cdot) \) by basis representation, PC or PLS with \( k_n \) elements.
   - Construct \( \epsilon_i^{**} = Y_i^* - \langle \mathcal{X}_i, \hat{\beta}^* \rangle \).
   - Compute \( \text{PCvM}_{n,p}^* = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \sum_{r=1}^n \epsilon_i^{**} \epsilon_j^{**} A_{ijr} \).
6. \( p \)-value \( \approx \# \{ \text{PCvM}_{n,p} \leq \text{PCvM}_{n,p}^* \} / B \).
The weak convergence of the process $R_n(u, \gamma)$ indexed in $\Pi = \mathbb{R} \times S_H$ is a very difficult task!

A related process is given by

$$\tilde{R}_n(u) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \left( Y_i - \langle X_i, \hat{\beta} \rangle \right) \mathbb{1}\{\langle x_i, \gamma \rangle \},$$

with a fixed but randomly chosen $\gamma \sim \omega$.

If $\omega$ is a non–degenerated Gaussian measure, then:

$$H_0 \iff \mathbb{E} [Y - \langle X, \beta \rangle | \langle X, \gamma \rangle] = 0, \text{ for some } \beta \in H.$$ 

In this way it is possible to obtain the weak convergence of $\tilde{R}_n$ indexed in $u \in \mathbb{R}$ and obtain statistical tests for testing $H_0$. 
Simulation scenario for the simple hypothesis

- Null Hypothesis: $H_0: m(\mathcal{X}) = \langle \mathcal{X}, \beta_0 \rangle$, where $\beta_0(t) = 0$, $t \in [0, 1]$.
- Deviations from the null
  - $H_{1,k}: \beta_{1,k}(t) = \gamma_k \cdot (t - 0.5)$, $k = 1, 2, 3$ with coefficients $\gamma_1 = 0.25$, $\gamma_2 = 0.65$ and $\gamma_3 = 1.00$.
  - $H_{2,k}: \beta_{2,k}(t) = \eta_k \cdot \sin(2\pi t^3)^3$, $k = 1, 2, 3$, with $\eta_1 = 0.10$, $\eta_2 = 0.20$ and $\eta_3 = 0.50$
  - $H_{3,k}: Y = \langle \mathcal{X}, \beta_0 \rangle + \delta_k \langle \mathcal{X}, \mathcal{X} \rangle + \varepsilon$
    $k = 1, 2, 3$ with $\delta_1 = 0.005$, $\delta_2 = 0.010$ and $\delta_3 = 0.015$
GOF tests for regression models: the functional data case

Simulation study

Simple hypothesis

Model 1: deviations from $\beta(t) = 0$

$\beta(t) = 0$

$\beta(t) = \beta_a(t)$

$\beta(t) = \beta_{a1}(t)$

$\beta(t) = \beta_{a2}(t)$

$\beta(t) = \beta_{a3}(t)$

Model 2: deviations from $\beta(t) = 0$

$\beta(t) = 0$

$\beta(t) = \beta_a(t)$

$\beta(t) = \beta_{a1}(t)$

$\beta(t) = \beta_{a2}(t)$

$\beta(t) = \beta_{a3}(t)$

Figure: Beta coefficients for the null and the alternatives.
Figure: Densities of the response $Y$ for the null and the alternatives. The two first deviations corresponds to the model $Y = \delta_i \langle \mathcal{X}, \beta_i \rangle + \varepsilon$, with $i = 1, 2$. The third deviation corresponds to $Y = \langle \mathcal{X}, \beta \rangle + \delta \langle \mathcal{X}, \mathcal{X} \rangle + \varepsilon$, with $\delta = 0.005, 0.010, 0.015$. 
Competing methods for the simple hypothesis

- Delsol et al. (2011) propose a test statistic for $H_0 : m(X) = m_0(X)$:

$$T_n = \frac{1}{n} \sum_{j=1}^{n} \left( \sum_{i=1}^{n} \epsilon_i K \left( \frac{d(X_j, X_i)}{h} \right) \right)^2 \omega(X_j),$$

where $K$ is a kernel function, $d$ is a semimetric ($L_2$), $h$ is the bandwidth (0.25, 0.50, 0.75 and 1.00) and $\omega$ a weight function (uniform).

- González-Manteiga et al. (2012) extends the ideas of the classical $F$–test to the functional framework, resulting a statistic to test the null hypothesis of no interaction inside the functional linear model:

$$D_n = \left\| \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}) (Y_i - \bar{Y}) \right\|_H.$$
## Simulation study

Simple hypothesis

<table>
<thead>
<tr>
<th>Models</th>
<th>$F$–test</th>
<th>PCvM test</th>
<th>Delsol test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$h = 0.25$</td>
</tr>
<tr>
<td>$H_0$</td>
<td>0.058</td>
<td>0.043</td>
<td>0.055</td>
</tr>
<tr>
<td>$H_{1,1}$</td>
<td>0.063</td>
<td>0.069</td>
<td>0.070</td>
</tr>
<tr>
<td>$H_{1,2}$</td>
<td>0.166</td>
<td>0.079</td>
<td>0.358</td>
</tr>
<tr>
<td>$H_{1,3}$</td>
<td>0.422</td>
<td>0.137</td>
<td>0.819</td>
</tr>
<tr>
<td>$H_{2,1}$</td>
<td>0.253</td>
<td>0.053</td>
<td>0.068</td>
</tr>
<tr>
<td>$H_{2,2}$</td>
<td>0.952</td>
<td>0.336</td>
<td>0.314</td>
</tr>
<tr>
<td>$H_{2,3}$</td>
<td>1.000</td>
<td>0.904</td>
<td>0.795</td>
</tr>
<tr>
<td>$H_{3,1}$</td>
<td>0.036</td>
<td>0.173</td>
<td>0.051</td>
</tr>
<tr>
<td>$H_{3,2}$</td>
<td>0.057</td>
<td>0.691</td>
<td>0.207</td>
</tr>
<tr>
<td>$H_{3,3}$</td>
<td>0.056</td>
<td>0.998</td>
<td>0.802</td>
</tr>
</tbody>
</table>

**Table:** Empirical power of the competing tests for the simple hypothesis $H_0 : m(X) = \langle X, \beta_0 \rangle$, $\beta_0 = 0$ and significance level $\alpha = 0.05$. Noise has a normal distribution with zero mean and standard deviation 0.10.
Simulation study

Simple hypothesis

<table>
<thead>
<tr>
<th>Models</th>
<th>$p = 1$</th>
<th>$p = 2$</th>
<th>$p = 3$</th>
<th>$p = 4$</th>
<th>$p = 5$</th>
<th>$p = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$</td>
<td>0.035</td>
<td>0.037</td>
<td>0.037</td>
<td>0.036</td>
<td>0.036</td>
<td>0.037</td>
</tr>
<tr>
<td>$H_{1,2}$</td>
<td>0.043</td>
<td>0.098</td>
<td>0.085</td>
<td>0.080</td>
<td>0.074</td>
<td>0.074</td>
</tr>
<tr>
<td>$H_{2,2}$</td>
<td>0.529</td>
<td>0.407</td>
<td>0.375</td>
<td>0.359</td>
<td>0.350</td>
<td>0.347</td>
</tr>
<tr>
<td>$H_{3,2}$</td>
<td>0.657</td>
<td>0.703</td>
<td>0.707</td>
<td>0.708</td>
<td>0.709</td>
<td>0.709</td>
</tr>
</tbody>
</table>

**Table:** Empirical power of the PCvM test for the simple hypothesis $H_0 : m(\mathcal{X}) = \langle \mathcal{X}, \beta_0 \rangle , \beta_0 = 0$, for different numbers $p$ of FPC. The significance level is $\alpha = 0.05$ and noise has a normal distribution with zero mean and standard deviation 0.10.
Simulation scenario for the composite hypothesis

- Null hypothesis: we have considered three different null models of the form

\[ H_{j,0} : \quad Y = \langle \mathcal{X}, \beta_j \rangle + \varepsilon, \]

with \( j = 1, 2, 3 \) and \( \beta_1(t) = \sin(2\pi t) - \cos(2\pi t) \), \( \beta_2(t) = t - (t - 0.75)^2 \) and \( \beta_3(t) = t + \cos(2\pi t) \), \( t \in [0, 1] \).

- Deviations from the linear regression model are considered introducing a second order term \( \langle \mathcal{X}, \mathcal{X} \rangle \) with three different weights \( (\delta_1 = 0.01, \delta_2 = 0.05 \text{ and } \delta_3 = 0.10) \), representing the alternatives \( H_{j,k} \):

\[ H_{j,k} : \quad Y = \langle \mathcal{X}, \beta_j \rangle + \delta_k \langle \mathcal{X}, \mathcal{X} \rangle + \varepsilon. \]
**Figure:** Beta coefficients and densities of the response $Y$ for the null and the alternatives.
Table: Empirical power of the PCvM test for the composite hypothesis

\[ H_0 : m(\mathcal{X}) = \langle \mathcal{X}, \beta \rangle \] and for three estimating methods of \( \beta \).

\( \text{Noise} \sim \mathcal{N}(0, 0.10^2) \).
Simulation study

Composite hypothesis

<table>
<thead>
<tr>
<th>Models</th>
<th>$p = 1$</th>
<th>$p = 2$</th>
<th>$p = 3$</th>
<th>$p = 4$</th>
<th>$p = 5$</th>
<th>$p = 6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_{1,0}$</td>
<td>0.044</td>
<td>0.053</td>
<td>0.049</td>
<td>0.056</td>
<td>0.060</td>
<td>0.062</td>
</tr>
<tr>
<td>$H_{1,1}$</td>
<td>0.054</td>
<td>0.062</td>
<td>0.079</td>
<td>0.079</td>
<td>0.085</td>
<td>0.089</td>
</tr>
<tr>
<td>$H_{1,2}$</td>
<td>0.192</td>
<td>0.410</td>
<td>0.685</td>
<td>0.743</td>
<td>0.755</td>
<td>0.757</td>
</tr>
<tr>
<td>$H_{1,3}$</td>
<td>0.577</td>
<td>0.911</td>
<td>0.996</td>
<td>0.997</td>
<td>0.997</td>
<td>0.997</td>
</tr>
</tbody>
</table>

Table: Empirical power of the PCvM test for the composite hypothesis $H_0 : m \in \{ \langle \cdot, \beta \rangle : \beta \in \mathbb{H} \}$, for different numbers $p$ of FPC. The significance level is $\alpha = 0.05$ and noise has a normal distribution with zero mean and standard deviation 0.10.
## Coefficient estimation

<table>
<thead>
<tr>
<th>Models</th>
<th>$n$</th>
<th>Optimal basis</th>
<th>FPC estimation</th>
<th>FPLS estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$\alpha=0.10$</td>
<td>$\alpha=0.05$</td>
<td>$\alpha=0.10$</td>
</tr>
<tr>
<td>$H_{1,0}$</td>
<td>50</td>
<td>0.140</td>
<td>0.116</td>
<td>0.123</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.119</td>
<td>0.102</td>
<td>0.104</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.111</td>
<td>0.092</td>
<td>0.107</td>
</tr>
<tr>
<td>$H_{1,1}$</td>
<td>50</td>
<td>0.157</td>
<td>0.098</td>
<td>0.139</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.160</td>
<td>0.118</td>
<td>0.151</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.212</td>
<td>0.149</td>
<td>0.209</td>
</tr>
<tr>
<td>$H_{1,2}$</td>
<td>50</td>
<td>0.607</td>
<td>0.351</td>
<td>0.551</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>0.845</td>
<td>0.418</td>
<td>0.809</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>0.982</td>
<td>0.537</td>
<td>0.978</td>
</tr>
<tr>
<td>$H_{1,3}$</td>
<td>50</td>
<td>0.957</td>
<td>0.505</td>
<td>0.933</td>
</tr>
<tr>
<td></td>
<td>100</td>
<td>1.000</td>
<td>0.474</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>200</td>
<td>1.000</td>
<td>0.549</td>
<td>1.000</td>
</tr>
</tbody>
</table>

**Table:** Empirical power of the PCvM test for the composite hypothesis $H_0 : m \in \{ \langle \cdot, \beta \rangle : \beta \in \mathbb{H} \}$ and for different sample sizes $n$. Noise $\sim \mathcal{N}(0, 0.10^2)$. 

---

UC3M 2013 – W. González-Manteiga et al.
**Figure:** AEMET temperatures and estimated functional coefficient $\hat{\beta}$. Response: average wind speed.

- $H_0 : Y = \langle X, \beta \rangle + \varepsilon$. $p$–value is 0.119, no rejection of FLM.
- Simple hypothesis $H_0 : \beta = 0$

<table>
<thead>
<tr>
<th></th>
<th>PCvM</th>
<th>F–test</th>
<th>Delsol <em>et al.</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$–value</td>
<td>0.062</td>
<td>0.002</td>
<td>0.000</td>
</tr>
</tbody>
</table>
Figure: Tecator dataset. Absorbance, and first derivative by fat content.

- $H_0 : Y = \langle \mathcal{X}, \beta \rangle + \varepsilon$. $p$–value is 0.0000, evidence for rejecting FLM.
- Simple hypothesis $H_0 : \beta = 0$, $p$–value: 0.0000, significative dependence between $Y$ and $\mathcal{X}$, but *not a linear one*.
- Same results for first and second derivative ($\mathcal{X}', \mathcal{X}''$).
Figure: SP500 index and estimated functional coefficient $\hat{\beta}$ with 3 PLS. Response: mean index for the next session.

- Idea: predict the mean of the tomorrow’s session ($Y_{t+1} = \bar{X}_{t+1}$) with the today’s curve, $X_t$.
- $H_0 : Y_{t+1} = \langle X_t, \beta \rangle + \varepsilon$. $p$–value is 0.5604, no rejection of FLM.
- $F$–test: $p$–value = 0.000.
Conclusions

- We have proposed a goodness–of–fit test for the functional linear model based on empirical process as a generalization of a previous test of Escanciano (2006) with results quite promising.

- The test is easy to compute and simple to calibrate using wild bootstrap and it is available through the R-package \texttt{fda.usc} of Febrero–Bande and Oviedo–De la Fuente (2011).

- Obvious extensions of this test can include other possible forms of the regression function (several covariates, non linear form, . . .)
Thanks for your attention!