Bootstrap Calibration in Functional Linear Regression Models with Applications

Wenceslao González-Manteiga
(jointly with Adela Martínez-Calvo)

Departamento de Estadística e I.O.
Universidad de Santiago de Compostela (Spain)

USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGACIÓN OPERATIVA

COMPSTAT'2010, Paris (France)
August 23, 2010

Outline

1. Introduction
   - Bootstrap in finite dimensional case
   - Bootstrap in functional case
2. Bootstrap calibration in functional linear models
   - FPCA-type estimates
   - Confidence intervals for prediction
   - Test for lack of dependence
   - Test for equality of linear models
3. Simulation study and real data application
   - Confidence intervals for prediction
   - Test for lack of dependence
   - Test for equality of linear models
   - Real data application
4. Conclusions

Outline

1. Introduction
   - Bootstrap in finite dimensional case
   - Bootstrap in functional case
2. Bootstrap calibration in functional linear models
   - FPCA-type estimates
   - Confidence intervals for prediction
   - Test for lack of dependence
   - Test for equality of linear models
3. Simulation study and real data application
   - Confidence intervals for prediction
   - Test for lack of dependence
   - Test for equality of linear models
   - Real data application
4. Conclusions

Aim

- Our work focuses on the **functional linear model with scalar response** given by
$$Y = \langle \theta, X \rangle + \epsilon,$$
where Y and ϵ are real r.v., X is a r.v. valued in a Hilbert space $\mathcal{H}$, and $\theta \in \mathcal{H}$ is the fixed model parameter.
- From an initial sample $\{(X_i, Y_i)\}_{i=1}^n$, a **bootstrap resampling** is proposed
$$Y_i^* = \langle \hat{\theta}, X_i \rangle + \hat{\epsilon}_i^*, \quad i = 1, \ldots, n$$
where $\hat{\theta}$ is a pilot estimator, and $\hat{\epsilon}_i^*$ is a bootstrap error.
- This procedure allows us to calibrate some interesting distributions and to test different hypotheses related with θ.

Bootstrap in finite dimensional case: first applications

- Since its introduction by Efron (1979), the bootstrap method resulted in a new distribution approximation applicable to a large number of situations as the calibration of pivotal quantities in the finite dimensional context (see Bickel and Freedman (1981) and Singh (1981)).

BICKEL, P.J. and FREEDMAN, D.A. (1981): Some asymptotic theory for the bootstrap. *Annals of Statistics 9, 1196-1217*.

EFRON, B. (1979): Bootstrap methods: another look at the jackknife. *Annals of Statistics 7, 1-26*.

SINGH, K. (1981): On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics 9, 1187-1195*.

Bootstrap in finite dimensional case: linear regression

$$Y = X^t\theta + \epsilon,$$

where Y and ϵ are univariate r.v., X is a p-dimensional r.v. $(p \leq n)$, and θ is a p-vector of unknow parameters.

Theorem (Freedman (1981); $\hat{\theta}$: least squares estimator)

Let us assume that $\mathbb{E}(\epsilon_i^2|X_i) = \sigma^2$ *where* $\sigma^2 = \mathbb{E}(\epsilon_i^2)$.

- $n^{1/2}(\hat{\theta} - \theta)$ *is asymptotically* $\mathcal{N}(0, \sigma^2 \left[\mathbb{E}(X^t X)\right]^{-1})$.
- *The conditional law of* $n^{1/2}(\hat{\theta}^* - \hat{\theta})$ *goes weakly to* $\mathcal{N}(0, \sigma^2 \left[\mathbb{E}(X^t X)\right]^{-1})$.

FREEDMAN, D.A. (1981): Bootstrapping regression models. *Annals of Statistics 9, 1218-1228.*

Bootstrap in finite dimensional case: nonparametric regression

$$Y = m(X) + \epsilon,$$

where Y and ϵ are univariate r.v., X is a p-dimensional r.v., and m is a unknown regression function.

Theorem (Cao-Abad (1991); $\hat{m}_h(\cdot)$: kernel estimator)

$$\sup_{y \in \mathbb{R}} \left| P_{XY}((nh^p)^{1/2}(\hat{m}_h^*(x) - \hat{m}_g(x)) \leq y) - P_X((nh^p)^{1/2}(\hat{m}_h(x) - m(x)) \leq y) \right| \xrightarrow{P} 0$$

where P_{XY} denotes the probability measure under the bootstrap resampling plan, and P_X denotes the probability conditionally on $\{X_i\}_{i=1}^n$.

CAO-ABAD, R. (1991): Rate of convergence for the wild bootstrap in nonparametric regression. *Annals of Statistics 19, 2226-2231*.

Bootstrap in functional case: first applications I

- Cuevas et al. (2004) developed a sort of parametric bootstrap to obtain quantiles for an anova test.
- Cuevas et al. (2006) proposed bootstrap confidence bands for several functional estimators as the sample functional mean or the trimmed functional mean.
- Hall and Vial (2006) studied the finite dimensionality of functional data using a bootstrap approximation.
- Bathia et al. (2010) used bootstrap to identify the dimensionality of curve time series.

Bootstrap in functional case: first applications II

- BATHIA, N., YAO, Q. and ZIEGELMANN, F. (2010): Identifying the finite dimensionality of curve time series. *Annals of Statistics (to appear)*.
- CUEVAS, A., FEBRERO, M. and FRAIMAN, R. (2004): An Anova test for functional data. *Computational Statistics & Data Analysis 47, 111-122*.
- CUEVAS, A., FEBRERO, M. and FRAIMAN, R. (2006): On the use of the bootstrap for estimating functions with functional data. *Computational Statistics & Data Analysis 51, 1063-1074*.
- HALL, P. and VIAL, C. (2006): Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society Series B 68, 689-705*.

Bootstrap in functional case: linear regression

$$Y = \langle \theta, X \rangle + \epsilon,$$

where Y and ϵ are univariate r.v., X is a functional r.v. valued in a Hilbert space $\mathcal{H}$, and $\theta \in \mathcal{H}$ is a functional unknown parameter.

Theorem (González-Manteiga and Martínez-Calvo (2010); $\hat{\theta}_c$: FPCA-type estimator)

$$\sup_{y \in \mathbb{R}} \left| P_{XY}(n^{1/2}(\langle \hat{\theta}^*_{c,d}, x \rangle - \langle \hat{\theta}_d, x \rangle) \leq y) - P_X(n^{1/2}(\langle \hat{\theta}_c, x \rangle - \langle \hat{\Pi}_{k_n^c} \theta, x \rangle) \leq y) \right| \xrightarrow{P} 0,$$

where $\hat{\Pi}_{k_n^c}$ is the projection on the first k_n^c eigenfunctions of Γ_n, P_{XY} denotes the probability conditionally on $\{(X_i, Y_i)\}_{i=1}^n$, and P_X denotes the probability conditionally on $\{X_i\}_{i=1}^n$.

GONZÁLEZ-MANTEIGA, W. and MARTÍNEZ-CALVO, A. (2010): Bootstrap in functional linear regression. *Journal of Statistical Planning and Inference (to appear)*.

Bootstrap in functional case: nonparametric regression

$$Y = m(X) + \epsilon,$$

where Y and ϵ are univariate r.v., X is a functional r.v., and m is a unknown regression function.

Theorem (Ferraty et al. (2010); $\hat{m}_h(\cdot)$: kernel estimator for functional case)

$$\sup_{y \in \mathbb{R}} \Big| P_{XY}((nF_x(h))^{1/2}(\hat{m}_h^*(x) - \hat{m}_g(x)) \leq y) - P((nF_x(h))^{1/2}(\hat{m}_h(x) - m(x)) \leq y) \Big| \xrightarrow{a.s.} 0$$

where P_{XY} denotes the probability conditionally on $\{X_i, Y_i\}_{i=1}^n$, and $F_x(\cdot)$ is the small ball probability given by $F_x(t) = P(X \in B(x,t))$.

FERRATY, F., VAN KEILEGOM, I. and VIEU, P (2010): On the validity of the bootstrap in non-parametric functional regression. *Scandinavian Journal of Statistics 37, 286-306*.

Bootstrap validity for regression models

X	Linear regression model
p-dimensional	$Y = X^t\theta + \epsilon$ $n^{1/2}(\hat{\theta}^* - \hat{\theta}) \leftrightarrow n^{1/2}(\hat{\theta} - \theta)$
functional	$Y = \langle \theta, X \rangle + \epsilon$ $n^{1/2}(\langle \hat{\theta}^*_{c,d}, x\rangle - \langle \hat{\theta}_d, x\rangle) \leftrightarrow n^{1/2}(\langle \hat{\theta}_c, x\rangle - \langle \hat{\Pi}_{k_n^c}\theta, x\rangle)$

X	Nonparametric regression model
p-dimensional	$Y = m(X) + \epsilon$ $(nh^p)^{1/2}(\hat{m}^*_h(x) - \hat{m}_g(x)) \leftrightarrow (nh^p)^{1/2}(\hat{m}_h(x) - m(x))$
functional	$Y = m(X) + \epsilon$ $(nF_x(h))^{1/2}(\hat{m}^*_h(x) - \hat{m}_g(x)) \leftrightarrow (nF_x(h))^{1/2}(\hat{m}_h(x) - m(x))$

Outline

1. Introduction
   - Bootstrap in finite dimensional case
   - Bootstrap in functional case
2. Bootstrap calibration in functional linear models
   - FPCA-type estimates
   - Confidence intervals for prediction
   - Test for lack of dependence
   - Test for equality of linear models
3. Simulation study and real data application
   - Confidence intervals for prediction
   - Test for lack of dependence
   - Test for equality of linear models
   - Real data application
4. Conclusions

Functional linear model with scalar response

We have considered the **functional linear regression model with scalar response** given by

$$Y = \langle \theta, X \rangle + \epsilon,$$

where

- Y is a real r.v.,
- X is a zero-mean r.v. valued in a real separable Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ such that $\mathbb{E}(\|X\|^4) < +\infty$ (being $\|\cdot\| = \langle \cdot, \cdot \rangle^{1/2}$),
- $\theta \in \mathcal{H}$ is the model parameter which verifies $\|\theta\|^2 < +\infty$, and
- ϵ is a real r.v. satisfying that $\mathbb{E}(\epsilon) = 0$, $\mathbb{E}(\epsilon^2) = \sigma^2 < +\infty$, and $\mathbb{E}(\epsilon X) = 0$.

FPCA-type estimates: construction of the estimator I

- Let us define the **second moment operator** Γ and the **cross second moment operator** Δ

  $$\Gamma(x) = \mathbb{E}(\langle X, x\rangle X), \quad \Delta(x) = \mathbb{E}(\langle X, x\rangle Y), \quad \forall x \in \mathcal{H}.$$

  Moreover, $\{(\lambda_j, v_j)\}_j$ will denote the eigenvalues and eigenfunctions of Γ, assuming that $\lambda_1 > \lambda_2 > \ldots > 0$.

- From a sample $\{(X_i, Y_i)\}_{i=1}^n$, we can derive their empirical counterparts

  $$\Gamma_n(x) = n^{-1}\sum_{i=1}^n \langle X_i, x\rangle X_i, \quad \Delta_n(x) = n^{-1}\sum_{i=1}^n \langle X_i, x\rangle Y_i, \quad \forall x \in \mathcal{H},$$

  being $\{(\hat{\lambda}_j, \hat{v}_j)\}_{j=1}^\infty$ the eigenelements of Γ_n $(\hat{\lambda}_1 > \hat{\lambda}_2 > \ldots)$.

FPCA-type estimates: construction of the estimator II

- If $\sum_{j=1}^{\infty} (\Delta(v_j)/\lambda_j)^2 < +\infty$ and $Ker(\Gamma) = \{0\}$, then

$$\min_{\beta \in \mathcal{H}} \mathbb{E}[(Y - \langle \beta, X \rangle)^2]$$

has an unique solution: $\theta = \sum_{j=1}^{\infty} \frac{\Delta(v_j)}{\lambda_j} v_j$.

- Cardot et al. (2007) proposed the next estimators family

$$\hat{\theta}_c = \sum_{j=1}^{n} f_n^c(\hat{\lambda}_j) \Delta_n(\hat{v}_j) \hat{v}_j,$$

where $c = c_n$ satisfies that $c \to 0$ and $0 < c < \lambda_1$, and $\{f_n^c : [c, +\infty) \to \mathbb{R}\}_n$ is a sequence of positive functions.

CARDOT, H., MAS, A. and SARDA, P. (2007): CLT in functional linear regression models. *Probability Theory and Related Fields 138, 325-361*.

FPCA-type estimates: examples I

Example 1. When $f_n(x) = x^{-1} 1_{\{x \geq c\}}$, the estimator $\hat{\theta}_c$ is asymptotically equivalent to the **standard FPCA estimator**

$$\hat{\theta}_{k_n} = \sum_{j=1}^{k_n} \frac{\Delta_n(\hat{v}_j)}{\hat{\lambda}_j} \hat{v}_j.$$

CAI, T.T. and HALL, P. (2006): Prediction in functional linear regression. *Annals of Statistics 34, 2159-2179*.

CARDOT, H., FERRATY, F. and SARDA, P. (2003b): Spline estimators for the functional linear model. *Statistica Sinica 13, 571-591*.

HALL, P. and HOROWITZ, J.L. (2007): Methodology and convergence rates for functional linear regression. *Annals of Statistics 35, 70-91*.

HALL, P. and HOSSEINI-NASAB, M. (2006): On properties of functional principal components analysis. *Journal of the Royal Statistical Society Series B 68, 109-126*.

FPCA-type estimates: examples II

Example 2. If $f_n(x) = (x + \alpha_n)^{-1} 1_{\{x \geq c\}}$ for α_n a sequence of positive parameters, the estimator $\hat{\theta}_c$ is asymptotically equivalent to the **ridge-type estimator** proposed by Martínez-Calvo (2008)

$$\hat{\theta}_{k_n}^{\alpha_n} = \sum_{j=1}^{k_n} \frac{\Delta_n(\hat{v}_j)}{\hat{\lambda}_j + \alpha_n} \hat{v}_j.$$

MARTÍNEZ-CALVO, A. (2008): Presmoothing in functional linear regression. In: S. Dabo-Niang and F. Ferray (Eds.): *Functional and Operatorial Statistics*. Physica-Verlag, Heidelberg, 223-229.

Confidence intervals for prediction

OBJECTIVE. We want to obtain pointwise confidence intervals for a certain confidence level α, that is, $I_{x,\alpha} \subset \mathbb{R}$ such that

$$P(\langle \theta, x \rangle \in I_{x,\alpha}) = 1 - \alpha$$

for a fixed $x \in H$.

Asymptotic confidence intervals

When θ (or x) is very well approximated by the projection on the subspace spanned by the first k_n^c eigenfunctions of Γ_n, the Central Limit Theorem shown by Cardot et al. (2007) allows us to evaluate the following **approximated asymptotic confidence intervals** for $\langle \theta, x \rangle$

$$I_{x,\alpha}^{asy} = \left[\langle \hat{\theta}_c, x \rangle - \frac{\hat{t}_{n,x}^c \hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2}, \langle \hat{\theta}_c, x \rangle + \frac{\hat{t}_{n,x}^c \hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2} \right],$$

with $\hat{t}_{n,x}^c = \sqrt{\sum_{j=1}^{k_n^c} \hat{\lambda}_j [f_n^c(\hat{\lambda}_j)]^2 \langle x, \hat{v}_j \rangle^2}$, $\hat{\sigma}^2$ a consistent estimate of σ^2, and z_α the quantile of order α of $Z \sim \mathcal{N}(0,1)$. [1]

CARDOT, H., MAS, A. and SARDA, P. (2007): CLT in functional linear regression models. *Probability Theory and Related Fields 138, 325-361.*

[1] $k_n^c = \sup\{j : \lambda_j + \delta_j/2 \geq c\}$ ($\delta_1 = \lambda_1 - \lambda_2$ and $\delta_j = \min(\lambda_{j-1} - \lambda_j, \lambda_j - \lambda_{j+1})$ if $j \neq 1$).

Bootstrap confidence intervals I

Step 1. Obtain a pilot estimator $\hat{\theta}_d = \sum_{j=1}^{n} f_n^d(\hat{\lambda}_j)\Delta_n(\hat{v}_j)\hat{v}_j$, and calculate the residuals $\hat{\epsilon}_i = Y_i - \langle \hat{\theta}_d, X_i \rangle$ for $i = 1, \ldots, n$.

Step 2. **(Naive)** Draw $\hat{\epsilon}_1^*, \ldots, \hat{\epsilon}_n^*$ i.i.d. r.v. from the cumulative distribution of $\{\hat{\epsilon}_i - \bar{\hat{\epsilon}}\}_{i=1}^n$, where $\bar{\hat{\epsilon}} = n^{-1}\sum_{i=1}^n \hat{\epsilon}_i$.
(Wild) For $i = 1, \ldots, n$, define $\hat{\epsilon}_i^* = \hat{\epsilon}_i V_i$, where $\{V_i\}_{i=1}^n$ are i.i.d. r.v., independent of $\{(X_i, Y_i)\}_{i=1}^n$, such that $\mathbb{E}(V_1) = 0$ and $\mathbb{E}(V_1^2) = 1$.

Step 3. Construct $Y_i^* = \langle \hat{\theta}_d, X_i \rangle + \hat{\epsilon}_i^*$, for $i = 1, \ldots, n$.

Step 4. Build $\hat{\theta}_{c,d}^* = \sum_{j=1}^{n} f_n^c(\hat{\lambda}_j)\Delta_n^*(\hat{v}_j)\hat{v}_j$, where Δ_n^* is defined as $\Delta_n^*(\cdot) = n^{-1}\sum_{i=1}^n \langle X_i, \cdot \rangle Y_i^*$.

Remark. For consistency results, we need that $c \leq d$, so the no of PC used for $\hat{\theta}_{c,d}^*$ is larger than the no of PC used for $\hat{\theta}_d$. In some way, we should *oversmooth* when we calculate the pilot estimator.

Bootstrap confidence intervals II

Theorem (González-Manteiga and Martínez-Calvo (2010))

Let $\hat{\Pi}_{k_n^c}$ *be the projection on the first* k_n^c *eigenfunctions of* Γ_n*. Under certain hypotheses, for both the naive and the wild bootstrap,*

$$\sup_{y\in\mathbb{R}} \left| P_{XY}(n^{1/2}(\langle \hat{\theta}^*_{c,d}, x\rangle - \langle \hat{\theta}_d, x\rangle) \leq y) - P_X(n^{1/2}(\langle \hat{\theta}_c, x\rangle - \langle \hat{\Pi}_{k_n^c}\theta, x\rangle) \leq y) \right| \xrightarrow{P} 0,$$

where P_{XY} *denotes the probability conditionally on* $\{(X_i, Y_i)\}_{i=1}^n$*, and* P_X *denotes the probability conditionally on* $\{X_i\}_{i=1}^n$*.*

GONZÁLEZ-MANTEIGA, W. and MARTÍNEZ-CALVO, A. (2010): Bootstrap in functional linear regression. *Journal of Statistical Planning and Inference (to appear)*.

Bootstrap confidence intervals III

The theorem before ensures that the α-quantiles $q_\alpha(x)$ of the distribution of the true error $(\langle\hat{\theta}_c, x\rangle - \langle\theta, x\rangle)$ can be aproximated by the bootstrap α-quantiles $q^*_\alpha(x)$ of $(\langle\hat{\theta}^*_{c,d}, x\rangle - \langle\hat{\theta}_d, x\rangle)$.
Then we can build the next **bootstrap confidence intervals** for $\langle\theta, x\rangle$

$$I^*_{x,\alpha} = \left[\langle\hat{\theta}_c, x\rangle - q^*_{1-\alpha/2}(x), \langle\hat{\theta}_c, x\rangle - q^*_{\alpha/2}(x)\right].$$

Test for lack of dependence

OBJECTIVE. We want to test the null hypothesis

$$H_0 : \theta = 0,$$

being the alternative $H_1 : \theta \neq 0$.

Test for lack of dependence: asymptotic approach I

- Cardot et al. (2003a) deduced that testing H_0 is equivalent to test

$$H_0' : \ \Delta = 0.$$

- They proposed as test statistic

$$T_{1,n} = k_n^{-1/2} \left(\hat{\sigma}^{-2} ||\sqrt{n} \Delta_n \hat{A}_n||^2 - k_n \right),$$

where $\hat{A}_n(\cdot) = \sum_{j=1}^{k_n} \hat{\lambda}_j^{-1/2} \langle \cdot, \hat{v}_j \rangle \hat{v}_j$ and $\hat{\sigma}^2$ is an estimator of σ^2.

- Let us remark that

$$T_{1,n} = \frac{1}{\sqrt{k_n}} \left(\frac{n}{\hat{\sigma}^2} \sum_{j=1}^{k_n} \frac{(\Delta_n(\hat{v}_j))^2}{\hat{\lambda}_j} - k_n \right)$$

Test for lack of dependence: asymptotic approach II

- Under H_0', $T_{1,n} \xrightarrow{d} \mathcal{N}(0,2)$.
- Hence, H_0' is rejected if $|T_{1,n}| > \sqrt{2} z_{1-\alpha/2}$ (z_α the α-quantile of a $\mathcal{N}(0,1)$), and accepted otherwise.

Remark. For functional response Y, see Kokoszka et al. (2008).

CARDOT, H., FERRATY, F., MAS, A. and SARDA, P (2003a): Testing hypothesis in the functional linear model. *Scandinavian Journal of Statistics 30, 241-255*.

KOKOSZKA, P., MASLOVA, I., SOJKA, J. and ZHU, L. (2008): Testing for lack of dependence in the functional linear model. *Canadian Journal of Statistics 36, 1-16*.

Test for lack of dependence: bootstrap approach I

- The null hypothesis H_0 is equivalent to

$$H_0'' : \; ||\theta|| = 0.$$

- We know that

$$||\theta||^2 = \left\| \sum_{j=1}^{\infty} \langle \theta, v_j \rangle v_j \right\|^2 = \sum_{j=1}^{\infty} \langle \theta, v_j \rangle^2 = \sum_{j=1}^{\infty} \left(\frac{\Delta(v_j)}{\lambda_j} \right)^2.$$

Therefore, we can use the statistic

$$T_{2,n} = \sum_{j=1}^{k_n} \left(\frac{\Delta_n(\hat{v}_j)}{\hat{\lambda}_j} \right)^2.$$

Test for lack of dependence: bootstrap approach II

Step 1. **(Naive)** Draw $\hat{Y}_1^*, \ldots, \hat{Y}_n^*$ i.i.d. random variables from the cumulative distribution of $\{Y_i - \bar{Y}\}_{i=1}^n$, where $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$.

(Wild) For $i = 1, \ldots, n$, define $\hat{Y}_i^* = Y_i V_i$, where $\{V_i\}_{i=1}^n$ are i.i.d. r.v., independent of $\{(X_i, Y_i)\}_{i=1}^n$, such that $\mathbb{E}(V_1) = 0$ and $\mathbb{E}(V_1^2) = 1$.

Step 2. Build $\Delta_n^*(\cdot) = n^{-1} \sum_{i=1}^n \langle X_i, \cdot \rangle Y_i^*$, for $i = 1, \ldots, n$.

- The distribution of $T_{2,n}$ can be approximated by the distribution of

$$T_{2,n}^* = \sum_{j=1}^{k_n} \left(\frac{\Delta_n^*(\hat{v}_j)}{\hat{\lambda}_j} \right)^2.$$

- H_0'' is accepted when $T_{2,n} < q_{1-\alpha}^*$ being q_α^* the α-quantile of $T_{2,n}^*$.

Test for equality of linear models

OBJECTIVE. Let us assume that we have two samples

$$Y_{1,i_1} = \langle \theta_1, X_{1,i_1} \rangle + \epsilon_{1,i_1}, \quad 1 \leq i_1 \leq n_1,$$
$$Y_{2,i_2} = \langle \theta_2, X_{2,i_2} \rangle + \epsilon_{2,i_2}, \quad 1 \leq i_2 \leq n_2,$$

We also suppose that X_1 and X_2 have the same covariance operator Γ ($\{(\lambda_j, v_j)\}_j$ denote the eigenvalues and eigenfunctions of Γ) and $Var(\epsilon^1) = Var(\epsilon^2) = \sigma^2$.
The aim is to test

$$H_0 : \ ||\theta_1 - \theta_2|| = 0,$$

against $H_1 : \ ||\theta_1 - \theta_2|| \neq 0$.

Test for equality: asymptotic approach I

- Horváth et al. (2009) proposed the following test statistic

$$\hat{\Lambda}_{1,k_n} = n_1 \left(1 + \frac{n_1}{n_2}\right)^{-1} (\hat{\mu}_1 - \hat{\mu}_2)^t (\hat{\mathbf{\Sigma}}_{k_n}^{-1})(\hat{\mu}_1 - \hat{\mu}_2),$$

where $\hat{\mu}_l = ((\mathbf{X}_l)^t \mathbf{X}_l)^{-1} (\mathbf{X}_l)^t \mathbf{Y}_l$ being $\mathbf{X}_l(i,j) = \langle X_{l,i}, v_j \rangle$ for $l \in \{1,2\}$, and $\hat{\mathbf{\Sigma}}_{k_n} = \hat{\sigma}^2 diag(\hat{\lambda}_1^{-1}, \ldots, \hat{\lambda}_{k_n}^{-1})$.

- Let us note that

$$\hat{\Lambda}_{1,k_n} = \frac{1}{\hat{\sigma}^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} \sum_{j=1}^{k_n} \frac{\left(\Delta_{1,n}(\hat{v}_j) - \Delta_{2,n}(\hat{v}_j)\right)^2}{\hat{\lambda}_j},$$

where $\Delta_{l,n}(x) = n_l^{-1} \sum_{i=1}^{n_l} \langle X_{l,i}, x \rangle Y_{l,i}$, and $\{(\hat{\lambda}_j, \hat{v}_j)\}_j$ are the eigenelements of $\Gamma_n(x) = (n_1 + n_2)^{-1} \sum_{l=1}^{2} \sum_{i=1}^{n_l} \langle X_{l,i}, x \rangle X_{l,i}$.

Test for equality: asymptotic approach II

- Under H_0, $\hat{\Lambda}_{1,k_n} \xrightarrow{d} \chi^2_{k_n}$.
- H_0 is rejected if $\hat{\Lambda}_{1,k_n} > q_{1-\alpha}$, with q_α the α-quantile of $\chi^2_{k_n}$, and accepted otherwise.

HORVÁTH, L., KOKOSZKA, P. and REIMHERR, M. (2009): Two sample inference in functional linear models. *Canadian Journal of Statistics 37, 571-591*.

Testing for equality: bootstrap approach I

- Let us remark that

$$\|\theta_1 - \theta_2\|^2 = \sum_{j=1}^{\infty} \left(\frac{(\Delta_1 - \Delta_2)(v_j)}{\lambda_j} \right)^2.$$

We are going to consider the next test statistic

$$\hat{\Lambda}_{2,k_n} = \sum_{j=1}^{k_n} \left(\frac{(\Delta_{1,n} - \Delta_{2,n})(\hat{v}_j)}{\hat{\lambda}_j} \right)^2.$$

Testing for equality: bootstrap approach II

Step 1. Obtain $\hat{\theta}_d = \sum_{j=1}^{n_1+n_2} f_n^d(\hat{\lambda}_j)\Delta_n(\hat{v}_j)\hat{v}_j$ where $\Delta_n(x) = (n_1+n_2)^{-1}\sum_{l=1}^{2}\sum_{i=1}^{n_l}\langle X_{l,i}, x\rangle Y_{l,i}$. Calculate the residuals $\hat{\epsilon}_{l,i} = Y_{l,i} - \langle\hat{\theta}_d, X_{l,i}\rangle$ for all $i = 1,\ldots,n_l$, for $l \in \{1,2\}$.

Step 2. **(Naive)** Draw $\hat{\epsilon}^*_{l,1},\ldots,\hat{\epsilon}^*_{l,n_l}$ i.i.d. random variables from the cumulative distribution of $\{\hat{\epsilon}_{l,i} - \bar{\hat{\epsilon}}_l\}_{i=1}^{n_l}$, where $\bar{\hat{\epsilon}}_l = {n_l}^{-1}\sum_{i=1}^{n_l}\hat{\epsilon}_{l,i}$, for $l \in \{1,2\}$.
(Wild) For $i = 1,\ldots,n_l$, define $\hat{\epsilon}^*_{l,i} = \hat{\epsilon}_{l,i}V_i$, where $\{V_i\}_{i=1}^{n_l}$ are i.i.d. r.v., independent of $\{(X_{l,i}, Y_{l,i})\}_{i=1}^{n_l}$, such that $\mathbb{E}(V_1) = 0$ and $\mathbb{E}(V_1^2) = 1$, for $l \in \{1,2\}$.

Step 3. Build $\Delta^*_{l,n}(x) = {n_l}^{-1}\sum_{i=1}^{n_l}\langle X_{l,i}, x\rangle Y^*_{l,i}$, where $Y^*_{l,i} = \langle\hat{\theta}_d, X_{l,i}\rangle + \hat{\epsilon}^*_{l,i}$, for all $i = 1,\ldots,n_l$, for $l \in \{1,2\}$.

Testing for equality: bootstrap approach III

- H_0 is accepted when $\hat{\Lambda}_{2,k_n} < q^*_{1-\alpha}$ with q^*_α the α-quantile of

$$\hat{\Lambda}^*_{2,k_n} = \sum_{j=1}^{k_n} \left(\frac{(\Delta^*_{1,n} - \Delta^*_{2,n})(\hat{v}_j)}{\hat{\lambda}_j} \right)^2.$$

Otherwise, H_0 is rejected.

Outline

1. Introduction
   - Bootstrap in finite dimensional case
   - Bootstrap in functional case
2. Bootstrap calibration in functional linear models
   - FPCA-type estimates
   - Confidence intervals for prediction
   - Test for lack of dependence
   - Test for equality of linear models
3. Simulation study and real data application
   - Confidence intervals for prediction
   - Test for lack of dependence
   - Test for equality of linear models
   - Real data application
4. Conclusions

Confidence intervals: simulation study I

- We have simulated $ns = 500$ samples, each being composed of $n \in \{50, 100\}$ observations from a functional linear model

$$Y = \langle \theta, X \rangle + \epsilon,$$

being X a Brownian motion and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with signal-to-noise ratio $r = \sigma / \sqrt{\mathbb{E}(\langle X, \theta \rangle^2)} = 0.2$.

- The model parameter is

$$\theta(t) = \sin(4\pi t), \quad t \in [0, 1],$$

and both X and θ were discretized to 100 design points.

- We have fixed six deterministic curves x

$$x_1 = \sin(\pi t/2),\ x_2 = \sin(3\pi t/2),\ x_3 = t,$$
$$x_4 = t^2,\ x_5 = 2|t - 0.5|,\ x_6 = 2I_{t>0.5}.$$

Confidence intervals: simulation study II

Asymptotic	$I^{asy}_{x,\alpha} = \left[\langle\hat{\theta}_c, x\rangle - \frac{\hat{t}^c_{n,x}\hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2}, \langle\hat{\theta}_c, x\rangle + \frac{\hat{t}^c_{n,x}\hat{\sigma}}{\sqrt{n}} z_{1-\alpha/2}\right]$
Bootstrap	$I^{*}_{x,\alpha} = \left[\langle\hat{\theta}_c, x\rangle - q^*_{1-\alpha/2}(x), \langle\hat{\theta}_c, x\rangle - q^*_{\alpha/2}(x)\right]$

- To select k_n, we have used GCV technique. $\alpha \in \{0.05, 0.10\}$.
- For asymptotic intervals the estimation for the true variance σ^2 is the residual sum of squares where k_n is chosen by GCV.
- For the bootstrap intervals, we have considered different pilot values $\{\hat{k}_n - 5, \ldots, \hat{k}_n + 2\}$, where $\hat{k}_n$ is the number of principal components selected by GCV. Moreover, 1000 bootstrap iterations were done and wild bootstrap was considered.

Confidence intervals: $n = 50$

α	CI	x_1	x_2	x_3	x_4	x_5	x_6
5%	$I_{x,\alpha}^{asy}$	8.8 (1.15)	9.0 (3.44)	10.6 (1.02)	13.2 (1.15)	19.8 (3.83)	23.0 (5.46)
	$I_{x,\alpha}^{*}$ $\hat{k}_n + 2$	10.6 (1.14)	10.8 (3.38)	11.4 (1.01)	13.4 (1.14)	14.4 (4.53)	17.0 (6.33)
	$I_{x,\alpha}^{*}$ $\hat{k}_n + 1$	10.4 (1.15)	10.4 (3.41)	12.0 (1.02)	13.2 (1.14)	15.6 (4.45)	19.6 (6.27)
	$I_{x,\alpha}^{*}$ $\hat{k}_n$	10.6 (1.15)	11.6 (3.43)	11.6 (1.02)	13.6 (1.15)	14.4 (4.41)	18.8 (6.23)
	$I_{x,\alpha}^{*}$ $\hat{k}_n - 1$	6.4 (1.36)	8.8 (4.04)	8.0 (1.21)	10.2 (1.37)	11.2 (4.97)	15.2 (7.11)
	$I_{x,\alpha}^{*}$ $\hat{k}_n - 2$	**5.4 (1.67)**	**5.4 (4.99)**	5.8 (1.48)	7.4 (1.67)	7.6 (5.95)	10.8 (8.69)
	$I_{x,\alpha}^{*}$ $\hat{k}_n - 3$	4.4 (2.11)	3.2 (6.33)	**4.6 (1.88)**	**5.8 (2.11)**	6.4 (7.33)	9.8(10.97)
	$I_{x,\alpha}^{*}$ $\hat{k}_n - 4$	3.2 (2.62)	2.2 (7.74)	3.8 (2.32)	**4.2 (2.59)**	**5.0 (8.75)**	7.2(13.59)
	$I_{x,\alpha}^{*}$ $\hat{k}_n - 5$	2.2 (2.96)	1.8 (8.80)	2.8 (2.63)	2.4 (2.92)	4.2 (9.69)	**5.4(15.63)**
10%	$I_{x,\alpha}^{asy}$	17.4 (0.97)	15.2 (2.89)	18.0 (0.86)	19.2 (0.96)	26.0 (3.21)	29.8 (4.58)
	$I_{x,\alpha}^{*}$ $\hat{k}_n + 2$	17.2 (0.96)	18.0 (2.87)	18.2 (0.86)	19.6 (0.97)	21.0 (3.77)	26.8 (5.29)
	$I_{x,\alpha}^{*}$ $\hat{k}_n + 1$	17.2 (0.97)	18.0 (2.88)	18.8 (0.86)	19.4 (0.97)	20.6 (3.70)	26.2 (5.21)
	$I_{x,\alpha}^{*}$ $\hat{k}_n$	17.4 (0.97)	17.6 (2.89)	18.4 (0.86)	19.2 (0.97)	21.8 (3.66)	27.6 (5.16)
	$I_{x,\alpha}^{*}$ $\hat{k}_n - 1$	12.6 (1.15)	12.0 (3.42)	13.8 (1.03)	14.4 (1.16)	18.6 (4.08)	20.8 (5.87)
	$I_{x,\alpha}^{*}$ $\hat{k}_n - 2$	**10.4 (1.41)**	**10.8 (4.22)**	**10.0 (1.26)**	12.4 (1.41)	14.8 (4.86)	18.2 (7.13)
	$I_{x,\alpha}^{*}$ $\hat{k}_n - 3$	6.6 (1.78)	5.8 (5.35)	6.6 (1.59)	**8.0 (1.78)**	**10.8 (5.92)**	13.6 (8.93)
	$I_{x,\alpha}^{*}$ $\hat{k}_n - 4$	5.6 (2.21)	4.6 (6.55)	5.4 (1.96)	5.6 (2.18)	8.0 (7.03)	**10.0(10.97)**
	$I_{x,\alpha}^{*}$ $\hat{k}_n - 5$	3.8 (2.51)	2.6 (7.44)	4.0 (2.22)	4.8 (2.46)	7.0 (7.71)	7.2(12.58)

Table: Empirical coverage rate (lenght$\times 10^2$) for $n = 50$.

Confidence intervals: $n = 100$

α	CI		x_1	x_2	x_3	x_4	x_5	x_6
5%	$I_{x,\alpha}^{asy}$		6.0 (0.83)	6.8 (2.47)	6.0 (0.74)	6.6 (0.83)	14.6 (2.89)	14.2 (4.13)
	$I_{x,\alpha}^{*}$	$\hat{k}_n + 2$	7.0 (0.82)	7.8 (2.44)	7.6 (0.73)	8.0 (0.84)	8.4 (3.42)	7.4 (4.84)
	$I_{x,\alpha}^{*}$	$\hat{k}_n + 1$	7.4 (0.83)	7.2 (2.44)	8.2 (0.74)	7.8 (0.83)	9.2 (3.37)	8.6 (4.77)
	$I_{x,\alpha}^{*}$	$\hat{k}_n$	7.2 (0.83)	7.6 (2.44)	7.8 (0.74)	8.0 (0.84)	9.0 (3.32)	9.4 (4.72)
	$I_{x,\alpha}^{*}$	$\hat{k}_n - 1$	6.0 (0.90)	6.8 (2.66)	6.2 (0.80)	6.2 (0.91)	8.2 (3.46)	8.8 (4.93)
	$I_{x,\alpha}^{*}$	$\hat{k}_n - 2$	**4.2 (1.09)**	**5.0 (3.20)**	**4.2 (0.97)**	**4.8 (1.09)**	7.4 (4.03)	7.8 (5.82)
	$I_{x,\alpha}^{*}$	$\hat{k}_n - 3$	1.8 (1.37)	3.0 (4.08)	2.8 (1.22)	3.6 (1.38)	5.8 (5.01)	**5.4 (7.41)**
	$I_{x,\alpha}^{*}$	$\hat{k}_n - 4$	2.2 (1.69)	2.6 (5.04)	1.6 (1.50)	2.4 (1.69)	**4.4 (5.96)**	4.4 (9.31)
	$I_{x,\alpha}^{*}$	$\hat{k}_n - 5$	1.4 (1.97)	2.2 (5.87)	1.2 (1.75)	1.4 (1.96)	3.4 (6.69)	3.0(10.94)
10%	$I_{x,\alpha}^{asy}$		13.4 (0.69)	12.8 (2.08)	13.0 (0.62)	15.0 (0.70)	22.0 (2.43)	22.6 (3.47)
	$I_{x,\alpha}^{*}$	$\hat{k}_n + 2$	14.2 (0.70)	13.4 (2.06)	14.4 (0.62)	15.2 (0.70)	14.2 (2.85)	16.0 (4.04)
	$I_{x,\alpha}^{*}$	$\hat{k}_n + 1$	14.6 (0.70)	14.0 (2.06)	14.8 (0.62)	15.8 (0.70)	16.4 (2.80)	18.2 (3.96)
	$I_{x,\alpha}^{*}$	$\hat{k}_n$	13.8 (0.70)	14.0 (2.06)	14.8 (0.62)	15.8 (0.70)	17.0 (2.76)	18.2 (3.91)
	$I_{x,\alpha}^{*}$	$\hat{k}_n - 1$	**10.8 (0.76)**	12.2 (2.25)	11.8 (0.68)	12.0 (0.76)	16.4 (2.86)	17.4 (4.06)
	$I_{x,\alpha}^{*}$	$\hat{k}_n - 2$	8.6 (0.92)	**10.0 (2.70)**	**8.4 (0.82)**	**9.0 (0.92)**	13.6 (3.31)	14.0 (4.78)
	$I_{x,\alpha}^{*}$	$\hat{k}_n - 3$	6.8 (1.16)	5.8 (3.45)	5.8 (1.03)	6.8 (1.16)	**10.6 (4.09)**	**10.2 (6.04)**
	$I_{x,\alpha}^{*}$	$\hat{k}_n - 4$	5.4 (1.43)	4.4 (4.25)	4.2 (1.27)	5.2 (1.42)	8.6 (4.82)	7.4 (7.50)
	$I_{x,\alpha}^{*}$	$\hat{k}_n - 5$	3.6 (1.66)	3.2 (4.96)	3.2 (1.47)	3.6 (1.65)	5.6 (5.38)	4.8 (8.76)

Table: Empirical coverage rate (lenght$\times 10^2$) for $n = 100$.

Lack of dependence: simulation study I

- We have simulated $ns = 500$ samples, each being composed of $n \in \{50, 100\}$ observations from a functional linear model

$$Y = \langle \theta, X \rangle + \epsilon,$$

being X a Brownian motion and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with signal-to-noise ratio $r = \sigma / \sqrt{\mathbb{E}(\langle X, \theta \rangle^2)} \in \{0.5, 1, 2\}$ (under H_0, $\sigma = 1$).

- We have considered two model parameters

$$\theta_0(t) = 0, \quad t \in [0, 1],$$
$$\theta_1(t) = \sin(2\pi t^3)^3, \quad t \in [0, 1].$$

Both X and θ were discretized to 100 design points.

Lack of dependence: simulation study II

Statistical test	Distribution
$T_{1,n} = \frac{1}{\sqrt{k_n}}\left(\frac{n}{\hat{\sigma}^2}\sum_{j=1}^{k_n}\frac{(\Delta_n(\hat{v}_j))^2}{\hat{\lambda}_j} - k_n\right)$	$\mathcal{N}(0,2)$ $T_{1,n}^{*(a)} = \frac{1}{\sqrt{k_n}}\left(\frac{n}{(\hat{\sigma}^*)^2}\sum_{j=1}^{k_n}\frac{(\Delta_n^*(\hat{v}_j))^2}{\hat{\lambda}_j} - k_n\right)$ $T_{1,n}^{*(b)} = \frac{1}{\sqrt{k_n}}\left(\frac{n}{\hat{\sigma}^2}\sum_{j=1}^{k_n}\frac{(\Delta_n^*(\hat{v}_j))^2}{\hat{\lambda}_j} - k_n\right)$
$T_{2,n} = \sum_{j=1}^{k_n}\left(\frac{\Delta_n(\hat{v}_j)}{\hat{\lambda}_j}\right)^2$	$T_{2,n}^* = \sum_{j=1}^{k_n}\left(\frac{\Delta_n^*(\hat{v}_j)}{\hat{\lambda}_j}\right)^2$

- $k_n \in \{1,\ldots,20\}$; $\alpha \in \{0.2, 0.1, 0.05, 0.01\}$
- For asymptotic test $\hat{\sigma}^2 = \frac{1}{tr(I_n - S)}\sum_{i=1}^{n}(Y_i - SY_i)^2$, where S is the *hat matrix* for the penalized B-splines estimator (B-splines with degree 4 and 20 equispaced knots; second derivative for the penalty; ρ selected by GCV).
- For bootstrap test, the wild bootstrap was considered, and 1000 bootstrap iterations were done.

Lack of dependence: level $(\theta_0(t) = 0)$ I

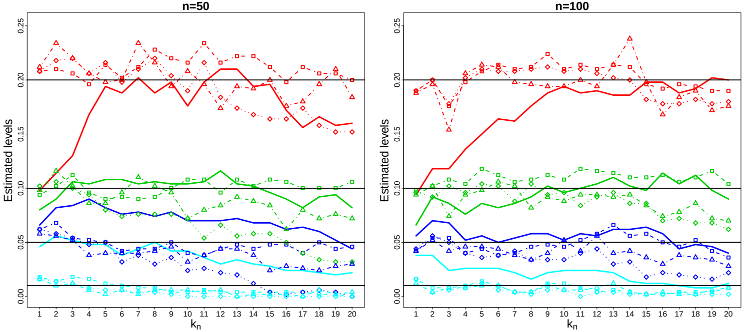


Figure: Estimated levels using the distribution of $\mathcal{N}(0,2)$ (solid line), $T^{*(a)}_{1,n}$ (square, dashed line), $T^{*(b)}_{1,n}$ (diamond, dotted line) and $T^{*}_{2,n}$ (triangle, dash-dotted line), for $\alpha = 0.2$ (red), 0.1 (green), 0.05 (blue) and 0.01 (light blue).

Lack of dependence: level $(\theta_0(t) = 0)$ II

		$\mathcal{N}(0,2)$			$T_{1,n}^{*(a)}$			$T_{1,n}^{*(b)}$			$T_{2,n}^{*}$		
n	α	$k_n = 5$	$k_n = 10$	$k_n = 20$	$k_n = 5$	$k_n = 10$	$k_n = 20$	$k_n = 5$	$k_n = 10$	$k_n = 20$	$k_n = 5$	$k_n = 10$	$k_n = 20$
50	20%	19.4	17.6	16.0	21.4	21.6	20.0	21.6	19.0	15.2	19.8	20.8	18.4
	10%	10.8	10.4	8.2	9.0	10.8	10.6	8.0	7.2	3.2	8.6	7.2	7.2
	5%	8.2	7.0	4.4	5.0	4.0	4.6	5.0	2.4	0.0	4.0	3.2	3.0
	1%	4.8	4.2	2.2	1.2	0.4	0.0	0.6	0.0	0.0	0.2	0.6	0.4
100	20%	15.0	19.4	20.0	20.8	21.0	19.0	21.0	20.8	18.0	21.4	19.4	17.6
	10%	8.6	9.6	9.0	11.8	10.8	10.4	10.4	9.6	6.2	9.8	8.8	7.0
	5%	5.6	5.2	4.0	4.4	4.6	3.6	3.6	3.4	2.2	4.6	5.2	2.8
	1%	2.6	2.4	1.2	1.4	1.2	0.8	1.2	0.6	0.2	1.0	0.6	0.8

Table: Comparison of the estimated levels (as percentage) for different values of k_n.

Lack of dependence: power ($\theta_1(t) = \sin(2\pi t^3)^3$) I

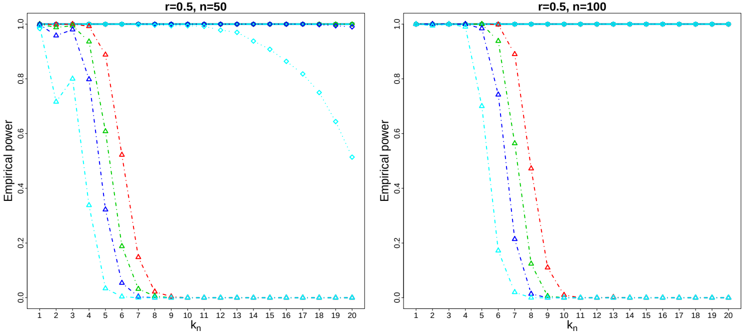


Figure: For $r = 0.5$, empirical power using the distribution of $\mathcal{N}(0,2)$ (solid line), $T_{1,n}^{*(a)}$ (square, dashed line), $T_{1,n}^{*(b)}$ (diamond, dotted line) and $T_{2,n}^{*}$ (triangle, dash-dotted line), for $\alpha = 0.2$ (red), 0.1 (green), 0.05 (blue) and 0.01 (light blue).

Lack of dependence: power ($\theta_1(t) = \sin(2\pi t^3)^3$) II

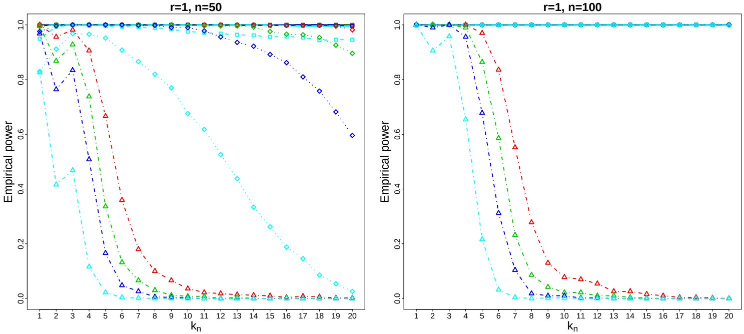


Figure: For $r = 1$, empirical power using the distribution of $\mathcal{N}(0,2)$ (solid line), $T^{*(a)}_{1,n}$ (square, dashed line), $T^{*(b)}_{1,n}$ (diamond, dotted line) and $T^{*}_{2,n}$ (triangle, dash-dotted line), for $\alpha = 0.2$ (red), 0.1 (green), 0.05 (blue) and 0.01 (light blue).

Lack of dependence: power ($\theta_1(t) = \sin(2\pi t^3)^3$) III

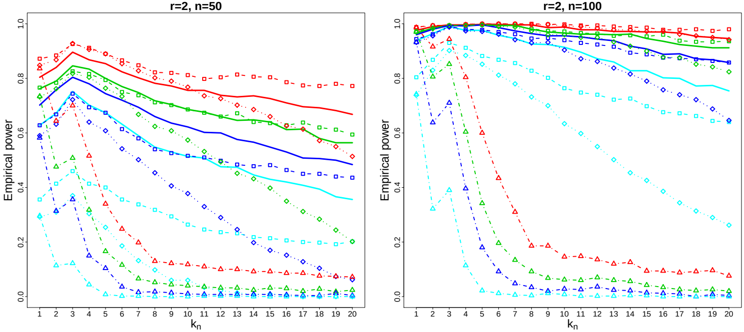


Figure: For $r = 2$, empirical power using the distribution of $\mathcal{N}(0, 2)$ (solid line), $T^{*(a)}_{1,n}$ (square, dashed line), $T^{*(b)}_{1,n}$ (diamond, dotted line) and $T^{*}_{2,n}$ (triangle, dash-dotted line), for $\alpha = 0.2$ (red), 0.1 (green), 0.05 (blue) and 0.01 (light blue).

Lack of dependence: power ($\theta_1(t) = \sin(2\pi t^3)^3$) IV

			$\mathcal{N}(0,2)$			$T^{*(a)}_{1,n}$			$T^{*(b)}_{1,n}$			$T^{*}_{2,n}$		
r	n	α	$k_n=5$	$k_n=10$	$k_n=20$	$k_n=5$	$k_n=10$	$k_n=20$	$k_n=5$	$k_n=10$	$k_n=20$	$k_n=5$	$k_n=10$	$k_n=20$
0.5	50	20%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	88.8	0.0	0.0
		10%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	60.8	0.0	0.0
		5%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.0	32.2	0.0	0.0
		1%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.4	51.4	3.4	0.0	0.0
	100	20%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	1.0	0.0
		10%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	0.0
		5%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.4	0.0	0.0
		1%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	70.0	0.0	0.0
1	50	20%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	98.2	66.6	3.6	0.2
		10%	100.0	100.0	100.0	100.0	100.0	99.8	100.0	99.8	89.6	33.6	0.8	0.0
		5%	100.0	100.0	99.8	100.0	100.0	99.6	100.0	99.0	59.6	16.6	0.2	0.0
		1%	100.0	100.0	99.6	99.6	97.6	94.6	95.2	67.6	2.6	2.2	0.0	0.0
	100	20%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.0	7.8	0.0
		10%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	86.4	2.2	0.0
		5%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	67.8	1.0	0.0
		1%	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.8	21.6	0.2	0.0
2	50	20%	85.4	75.6	66.8	89.0	81.2	77.2	89.0	76.8	51.4	34.0	11.8	7.2
		10%	80.0	68.6	56.4	79.4	68.6	59.4	76.4	57.4	20.2	16.6	4.0	2.4
		5%	74.4	62.2	48.4	67.4	51.6	43.6	60.8	37.8	6.2	10.4	1.0	0.4
		1%	67.4	51.4	35.6	40.0	26.4	20.2	25.4	6.0	0.0	0.8	0.0	0.0
	100	20%	99.8	98.8	94.6	100.0	99.8	98.0	100.0	99.2	94.2	60.0	14.6	7.6
		10%	99.6	96.6	91.2	99.6	97.2	93.6	99.6	96.0	82.4	34.2	6.2	2.0
		5%	99.6	95.6	85.8	97.8	94.0	85.8	97.2	90.4	64.6	18.0	2.8	0.4
		1%	97.6	91.4	75.4	88.2	76.4	64.0	85.2	63.4	26.2	2.2	0.8	0.0

Table: Comparison of the empirical power (as percentage) for different values of k_n and sample sizes.

Equality of linear models: simulation study I

- We have simulated $ns = 500$ pairs of samples, each being composed of $n_1, n_2 \in \{50, 100\}$ observations from the functional linear models

$$\begin{aligned} Y_{1,i_1} &= \langle \theta_1, X_{1,i_1} \rangle + \epsilon_{1,i_1}, \quad 1 \leq i_1 \leq n_1, \\ Y_{2,i_2} &= \langle \theta_2, X_{2,i_2} \rangle + \epsilon_{2,i_2}, \quad 1 \leq i_2 \leq n_2, \end{aligned}$$

  being X a Brownian motion and $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with signal-to-noise ratio $r = \sigma / \sqrt{\mathbb{E}(\langle X, \theta \rangle^2)} \in \{0.2\}$.

- We have considered the following model parameters

$$\begin{aligned} \theta_1(t) &= 2\sin(0.5\pi t) + 4\sin(1.5\pi t) + 5\sin(2.5\pi t), \quad t \in [0,1], \\ \theta_2(t) &= c\,(2\sin(0.5\pi t) + 4\sin(1.5\pi t) + 5\sin(2.5\pi t)), \quad t \in [0,1], \end{aligned}$$

  with $c \in \{1, 2\}$. Both X and θ were discretized to 100 points.

Equality of linear models: simulation study II

Statistical test	Distribution
$\hat{\Lambda}_{1,k_n} = \frac{1}{\hat{\sigma}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \sum_{j=1}^{k_n} \frac{(\Delta_{1,n}(\hat{v}_j)-\Delta_{2,n}(\hat{v}_j))^2}{\hat{\lambda}_j}$	$\chi^2_{k_n}$
	$\hat{\Lambda}^{*(a)}_{1,k_n} = \frac{1}{(\hat{\sigma}^*)^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \sum_{j=1}^{k_n} \frac{\left(\Delta^*_{1,n}(\hat{v}_j)-\Delta^*_{2,n}(\hat{v}_j)\right)^2}{\hat{\lambda}_j}$
	$\hat{\Lambda}^{*(b)}_{1,k_n} = \frac{1}{\hat{\sigma}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)} \sum_{j=1}^{k_n} \frac{\left(\Delta^*_{1,n}(\hat{v}_j)-\Delta^*_{2,n}(\hat{v}_j)\right)^2}{\hat{\lambda}_j}$
$\hat{\Lambda}_{2,k_n} = \sum_{j=1}^{k_n} \left(\frac{(\Delta_{1,n}-\Delta_{2,n})(\hat{v}_j)}{\hat{\lambda}_j}\right)^2$	$\hat{\Lambda}^*_{2,k_n} = \sum_{j=1}^{k_n} \left(\frac{(\Delta^*_{1,n}-\Delta^*_{2,n})(\hat{v}_j)}{\hat{\lambda}_j}\right)^2$

- $k_n \in \{1,\ldots,10\}$; $\alpha \in \{0.2, 0.1, 0.05, 0.01\}$
- For asymptotic test, $\hat{\sigma}^2$ is the residual standard deviation.
- For bootstrap test, the wild bootstrap was considered, and 1000 bootstrap iterations were done.

Equality of linear models: level $(c = 1)$ I

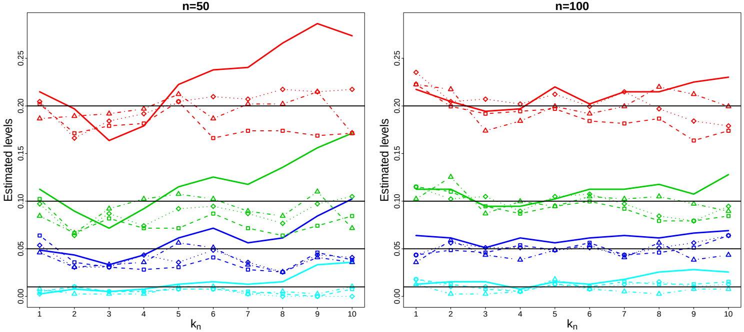


Figure: Estimated levels using the distribution of $\chi^2_{k_n}$ (solid line), $T^{*(a)}_{1,n}$ (square, dashed line), $T^{*(b)}_{1,n}$ (diamond, dotted line) and $T^{*}_{2,n}$ (triangle, dash-dotted line), for $\alpha = 0.2$ (red), 0.1 (green), 0.05 (blue) and 0.01 (light blue).

Equality of linear models: level ($c = 1$) II

		$\chi^2_{k_n}$			$T^{*(a)}_{1,n}$			$T^{*(b)}_{1,n}$			$T^{*}_{2,n}$		
n	α	$k_n=1$	$k_n=5$	$k_n=10$	$k_n=1$	$k_n=5$	$k_n=10$	$k_n=1$	$k_n=5$	$k_n=10$	$k_n=1$	$k_n=5$	$k_n=10$
50	20%	21.5	22.3	27.4	20.2	20.5	17.1	20.5	20.5	21.7	18.7	21.2	17.1
	10%	11.3	11.5	17.1	10.2	7.2	8.4	9.7	9.2	10.5	8.4	10.7	7.2
	5%	4.9	6.1	10.2	6.4	3.1	3.8	5.4	3.6	4.1	4.6	5.6	3.6
	1%	0.3	1.3	3.6	0.5	0.8	0.8	0.3	0.8	0.0	0.8	1.0	1.0
100	20%	21.7	22.0	23.0	22.3	19.7	17.4	23.5	21.2	17.9	22.3	19.9	19.9
	10%	11.3	10.2	12.8	11.5	9.5	8.4	11.5	10.5	9.5	10.2	9.5	9.0
	5%	6.4	5.6	6.9	4.3	4.9	6.4	4.3	4.9	6.4	3.6	4.9	4.3
	1%	1.3	1.5	2.6	1.8	1.3	1.5	1.8	1.3	1.3	1.3	1.8	0.8

Table: Comparison of the estimated levels (as percentage) for different values of k_n.

Equality of linear models: power ($c = 2$) I

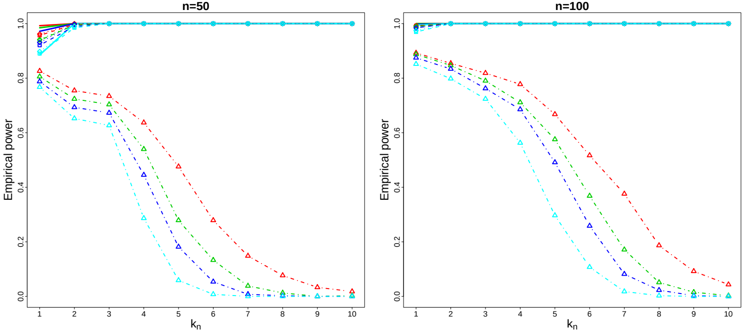


Figure: Empirical power using the distribution of $\chi^2_{k_n}$ (solid line), $T^{*(a)}_{1,n}$ (square, dashed line), $T^{*(b)}_{1,n}$ (diamond, dotted line) and $T^{*}_{2,n}$ (triangle, dash-dotted line), for $\alpha = 0.2$ (red), 0.1 (green), 0.05 (blue) and 0.01 (light blue).

Equality of linear models: power ($c = 2$) II

		$\chi^2_{k_n}$			$T^{*(a)}_{1,n}$			$T^{*(b)}_{1,n}$			$T^{*}_{2,n}$		
n	α	$k_n = 1$	$k_n = 5$	$k_n = 10$	$k_n = 1$	$k_n = 5$	$k_n = 10$	$k_n = 1$	$k_n = 5$	$k_n = 10$	$k_n = 1$	$k_n = 5$	$k_n = 10$
50	20.0	99.2	100.0	100.0	95.7	100.0	100.0	96.2	100.0	100.0	82.6	47.6	1.8
	10%	98.5	100.0	100.0	93.9	100.0	100.0	94.1	100.0	100.0	80.6	27.9	0.3
	5%	97.2	100.0	100.0	92.1	100.0	100.0	93.1	100.0	100.0	78.8	18.2	0.0
	1%	88.5	100.0	100.0	89.0	100.0	100.0	89.8	100.0	100.0	76.7	5.9	0.0
100	20%	100.0	100.0	100.0	99.2	100.0	100.0	99.2	100.0	100.0	89.3	66.8	4.3
	10%	100.0	100.0	100.0	99.0	100.0	100.0	99.0	100.0	100.0	88.7	57.5	0.3
	5%	99.7	100.0	100.0	98.5	100.0	100.0	98.5	100.0	100.0	87.5	49.1	0.0
	1%	99.5	100.0	100.0	96.9	100.0	100.0	97.4	100.0	100.0	85.2	29.7	0.0

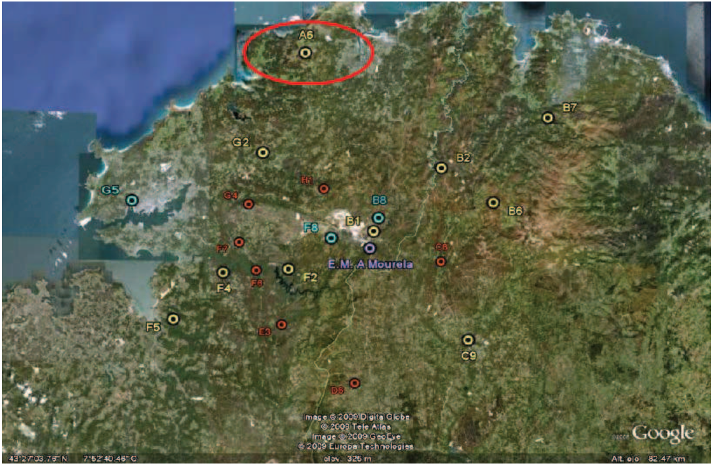
Table: Comparison of the empirical power (as percentage) for different values of k_n and sample sizes.

Real data application: atmospheric pollution data I

We are going to apply the tests exposed before to an environmental example.

- We have obtained concentrations of hourly averaged NO_x in the neighbourhood of a power station belongs to ENDESA, located in As Pontes in the Northwest of Spain. During unfavorable meteorological conditions, NO_x levels can quickly rise and cause an air-quality episode.
- The aim is to forecast NO_x with half an hour horizon to allow the power plant staff to avoid NO_x concentrations reaching the limit values fixed by the current environmental legislation.
- We have built a sample where each curve X corresponds to 240 consecutive minutal values of hourly averaged NO_x concentration, and the response Y corresponds to the NO_x value half an hour ahead (from Jan 2007 to Dec 2009).

Real data application: atmospheric pollution data II



Real data application: atmospheric pollution data III

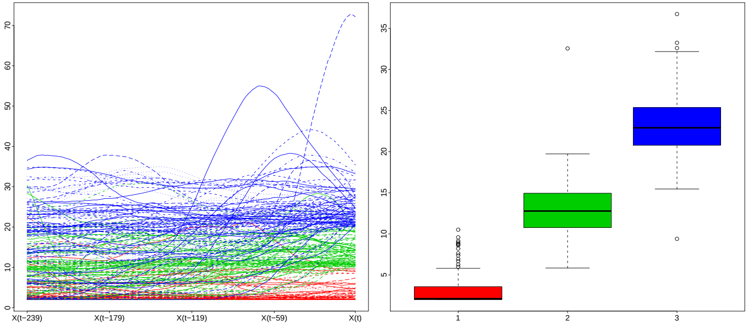


Figure: The curves X correspond to 240 consecutive minutal values of hourly averaged NO_x concentration (left), and the response Y corresponds to the NO_x value half an hour ahead (right). The data are classified in 3 bins depending on $X[240]$ value: < 10 (red), $10 - 20$ (green), and > 20 (blue).

Real data application: atmospheric pollution data IV

- Testing lack of dependence: $H_0 : \theta = 0$.

	$\chi^2_{k_n}$	$T^{*(a)}_{1,n}$	$T^{*(b)}_{1,n}$	$T^*_{2,n}$
$k_n = 5$	0	0	0	0.000
$k_n = 10$	0	0	0	0.002
$k_n = 20$	0	0	0	0.011

Table: P-values for testing the lack of dependence.

Real data application: atmospheric pollution data V

- Testing for equality of linear models: $H_0 : ||\theta_1 - \theta_2|| = 0$.

	Bin 1 & 2				Bin 1 & 3				Bin 2 & 3			
	$\chi^2_{k_n}$	$T^{*(a)}_{1,n}$	$T^{*(b)}_{1,n}$	$T^*_{2,n}$	$\chi^2_{k_n}$	$T^{*(a)}_{1,n}$	$T^{*(b)}_{1,n}$	$T^*_{2,n}$	$\chi^2_{k_n}$	$T^{*(a)}_{1,n}$	$T^{*(b)}_{1,n}$	$T^*_{2,n}$
$k_n = 5$	0.000	0.069	0.044	0.285	0	0.011	0.021	0.366	0.018	0.902	0.917	0.934
$k_n = 10$	0.001	0.954	0.931	0.461	0	0.012	0.009	0.807	0.000	0.458	0.302	0.748
$k_n = 20$	0.000	0.228	0.114	0.294	0	0.178	0.132	0.138	0.000	0.015	0.013	0.644

Table: P-values for testing equality between the bin 1 and the bin 2 (left), the bin 1 and the bin 3 (center), and the bin 2 and the bin 3(right).

Outline

1. Introduction
   - Bootstrap in finite dimensional case
   - Bootstrap in functional case
2. Bootstrap calibration in functional linear models
   - FPCA-type estimates
   - Confidence intervals for prediction
   - Test for lack of dependence
   - Test for equality of linear models
3. Simulation study and real data application
   - Confidence intervals for prediction
   - Test for lack of dependence
   - Test for equality of linear models
   - Real data application
4. Conclusions

Conclusions

- The proposed bootstrap methods seems to give test levels closer nominal ones than the tests based on the asymptotic distributions.
- In terms of the power of the tests, the statistic tests which include the error variance σ^2 are powerful that the tests which don't take it into account.
- In all the cases, the adequate k_n choice is quite important. This is still an open question.
- Further research: extension to functional response.

Bootstrap Calibration in Functional Linear Regression Models with Applications

Wenceslao González-Manteiga
(jointly with Adela Martínez-Calvo)

Departamento de Estadística e I.O.
Universidad de Santiago de Compostela (Spain)

COMPSTAT'2010, Paris (France)
August 23, 2010