

Proyecciones en Direcciones Aleatorias

o “Cómo reducir dimensionalidad sin pensar demasiado”

Gonzalo Barrera Borla

Análisis de Datos Funcionales, 2020

Universidad de Buenos Aires

1. Motivación

El problema: caracterizar e. a. ∞ – dimensionales

Una herramienta más: direcciones aleatorias

2. Algunas aplicaciones

Noción de Profundidad

Test de hipótesis vía proyecciones:

Bondad de ajuste

MLF con respuesta escalar

3. ANOVA funcional

Descripción del test

Consideraciones prácticas

En R: usando `fda.usc::fanova.RPm`

Motivación

El problema: caracterizar e. a. ∞ – dimensionales

Caracterizar elementos aleatorios en espacios ∞ -dimensionales implica expresarlos (i.e. proyectarlos) en *algún* subespacio de \mathbb{H} :

- bases fijas: Fourier, b-splines, ...
- descomposiciones data-dependientes: FPC/KLT, FCC, ...
- o en direcciones aleatorias.

¿Guarda alguna información útil una proyección aleatoria?

Resulta que sí, y bastante.

Una herramienta más: direcciones aleatorias

Sean

- X, Y elementos aleatorios (e. a.) en el espacio de Hilbert \mathbb{H} , $X \sim P$, $Y \sim Q$,
- $P \in DM(\mathbb{H})$, una distribución *determinada por sus momentos*, y
- \mathbb{E} el cono cerrado $\mathbb{E}(X, Y) := \{\alpha \in \mathbb{H} : \langle X, \alpha \rangle \sim \langle Y, \alpha \rangle\}$

Teorema Cuesta - Fraiman - Ransford

(Adaptado de [Cuesta 2007]) Sea μ una distribución Gaussiana no-degenerada en \mathbb{H} . Luego,

$$X \sim Y \iff \mu[\mathbb{E}(X, Y)] > 0$$

Aplicaciones

Profundidad aleatoria proyectada

Existen distintas nociones de profundidad posibles en \mathbb{H} :

- profundidad *integrada* de Fraiman y Muniz,
- el método de la *h-moda*,
- en direcciones aleatorias.

Profundidad aleatoria proyectada

Dada una muestra $\mathbf{X} = \{X_i\}_{i=1}^n$, sean

- $\alpha \in \mathbb{H}$ una dirección aleatoria independiente de \mathbf{X} y
- $\mathbf{x} = \{x_i : x_i = \langle X_i, \alpha \rangle\}_{i=1}^n$ la proyección de cada elemento de \mathbf{X} en ella.

La *profundidad aleatoria proyectada* muestral del dato original $RPD(X_i)$ se define como la profundidad unidimensional de x_i , expresada en términos del cuantil *aleatorio* de x_i en \mathbf{x} .

¡Ojo!

Al ser una medida aleatoria, hay que controlar su variabilidad, típicamente promediando sobre múltiples proyecciones $\{\alpha_i\}_{i=1}^n$.

Estas medidas de profundidad se pueden usadas como covariables en tareas de regresión [Cuevas 2007] y y clasificación [Cuevas 2007, Cuesta 2017].

Bondad de ajuste para familias paramétricas

Consideremos la hipótesis de pertenencia a cierta familia de distribuciones paramétricas

$$H_0 : P \in \mathcal{P} := \{P(\cdot, \theta) : \theta \in \Theta\} \quad \text{vs.} \quad H_1 : P \notin \mathcal{P}$$

¿Cómo se usaría el teorema de Cuesta-Fraiman-Ransford para construir tests a partir de proyecciones aleatorias?

Bondad de ajuste para familias paramétricas

[Cuesta 2007] considera familias paramétricas

1. invariantes

1.1 por locación (l -invariantes): $X \sim P_X \in \mathcal{P}, Y = X + m \Rightarrow P_Y \in \mathcal{P}$

1.2 por escala (s -invariantes): $X \sim P_X \in \mathcal{P}, Y = \Sigma X \Rightarrow P_Y \in \mathcal{P}$

1.3 elípticas (l - y s - invariantes)

2. ciertas familias no-invariantes, donde

- $\mathcal{P} := \{\mathcal{L}(sX + g) : s \in \mathbb{R} \wedge g \in V_n\}$ y
- V_n es un subespacio finito-dimensional de \mathbb{H}

y prueba que familias del tipo

1. están *determinadas en probabilidad* (d. p.) por una única proyección aleatoria.
2. con n parámetros están d. p. por $k = n + 1$ proyecciones aleatorias.

Bondad de ajuste para familias paramétricas

En cualquier caso, si la familia determinada por $\{h_i\}_{i=1}^k$ proyecciones aleatorias μ – generadas, resulta que la *hipótesis proyectada*

$$H_0^h : (P_{h_1}, \dots, P_{h_k}) \in \mathcal{P}_{h_1, \dots, h_k} := \{(P_{h_1}(\cdot, \theta), \dots, P_{h_k}(\cdot, \theta)) : \theta \in \Theta\}$$

es μ^k – casi seguramente (μ^k – c.s.) equivalente a la hipótesis original, ya que μ^k – c.s.

$$\begin{aligned}(P_{h_1}, \dots, P_{h_k}) \in \mathcal{P}_{h_1, \dots, h_k} &\iff P \in \mathcal{P} \\ H_0^h &\iff H_0\end{aligned}$$

Y para alguna medida apropiada de distancia d que mida el desvío de P c.r.a. \mathcal{P} bajo H_0 , μ^k – c.s.

$$\max_{i=1, \dots, k} d(P_{h_i}, P_{h_i}(\cdot, \theta)) = 0$$

Bondad de ajuste para familias paramétricas

Tomando por d la distancia Kolmogoro-Smirnov, podemos basar el test en el estadístico

$$D_n := \max_{i=1, \dots, k} \sqrt{nd} \left((\mathbb{P}_n)_{h_i}, P_{h_i}(\cdot, \theta) \right)$$

i.e. el máximo de k estadísticos K-S univariados, donde \mathbb{P}_n es la distribución empírica basada en $\{X_i\}_{i=1}^n$.

Restan dos dificultades:

- Cuando $k = 1$, el estadístico K-S es de distribución libre, pero para $k > 1$ ya no.
- Salvo para hipótesis “simples” $H_0 : P_0 = P(\cdot, \theta_0)$, es necesario estimar θ a partir de \mathbf{X} .

Así que tendremos que contentarnos con un estimador \hat{D}_n de D_n . Para aproximar la distribución de \hat{D}_n , sugieren un procedimiento bootstrap y prueban que vale bajo ciertas condiciones.

MLF con respuesta escalar

[Garcia 2014, Cuesta 2019] aplican estas ideas al modelo lineal funcional con respuesta escalar

$$Y = \langle X, \beta \rangle + \varepsilon = \int X(t) \beta(t) dt + \varepsilon$$

donde $X, \beta \in \mathbb{H}$, $Y \in \mathbb{R}$ y se sostienen los supuestos clásicos sobre ε detallados en los apuntes de clase.

La predicción de Y se hace a partir de la esperanza condicional

$$m(X) = E[Y|X] = \langle X, \beta \rangle$$

de manera que las siguientes expresiones son equivalentes:

- El MLF representa adecuadamente la relación entre X e Y
- $m(X) \in \mathcal{M} := \{\langle \cdot, \beta \rangle : \beta \in \mathbb{H}\}$

A partir de aquí, se pueden utilizar las técnicas antes mencionadas para construir tests

1. para hipótesis simples : $\beta = \beta_0$
2. incluido el caso particular de “no interacción”, $\beta_0(t) = 0 \forall t$, y
3. para la significatividad global, estimando previamente β

La clave, está en el siguiente lema, adaptado de [Patilea 2012], que permite caracterizar $H_0 : m \in \mathcal{M}$.

MLF con respuesta escalar

Consideremos

- $\mathbb{H} = L^2[a, b]$
- $\{\Psi_j\}_{j=1}^{\infty}$ una base de \mathbb{H} , no necesariamente ortogonal,
- la *esfera funcional* $\mathbb{S}_{\mathbb{H}} = \{h \in \mathbb{H} : \|h\|_{\mathbb{H}} = 1\}$, y
- la *esfera funcional p-dimensional*
 $\mathbb{S}_{\mathbb{H}}^p = \left\{ h = \sum_{j=1}^p x_j \Psi_j \in \mathbb{H} : \|h\|_{\mathbb{H}} = 1 \right\}.$

Lema [Patilea 2012]

Sea β un elemento de \mathbb{H} . Las siguientes afirmaciones son equivalentes:

1. $m(X) = \langle X, \beta \rangle \forall X \in \mathbb{H}$
2. $\mathbb{E}[Y - \langle X, \beta \rangle | X = x] = 0$ para casi todo (p.c.t.) $x \in \mathbb{H}$
3. $\mathbb{E}[Y - \langle X, \beta \rangle | \langle X, \gamma \rangle = u] = 0$ p.c.t. $u \in \mathbb{R}$ y $\forall \gamma \in \mathbb{S}_{\mathbb{H}}$
4. $\mathbb{E}[Y - \langle X, \beta \rangle | X = x] = 0$ p.c.t. $u \in \mathbb{R}$ y $\forall \gamma \in \mathbb{S}_{\mathbb{H}}^p, \forall p \geq 1$

ANOVA funcional

[Cuesta 2010] propone un procedimiento relativamente sencillo para diseños con dos factores, interacciones y covariables funcionales.

Sean

- \mathbb{H} un espacio de Hilbert separable con producto interno $\langle \cdot, \cdot \rangle$ medido w.l.g. en el intervalo $[0, 1]$, y
- $R, S \in \mathbb{N} : \forall (r, s) \in \{1, \dots, R\} \times \{1, \dots, S\}$

Y existen $X_i^{r,s}, i \in \{1, \dots, n_{r,s}\}$ e. a. en \mathbb{H} tales que

$$X_i^{r,s}(t) = m(t) + f(t) + g^s(t) + h^{r,s}(t) + \gamma(t) Y_i^{r,s} + \epsilon_i^{r,s}(t), \quad t \in [0, 1]$$

Descripción del test

$$X_i^{r,s}(t) = m(t) + f(t) + g^s(t) + h^{r,s}(t) + \gamma(t) Y_i^{r,s} + \epsilon_i^{r,s}(t), \quad t \in [0, 1]$$

donde

1. la función m es fija y describe la forma general del proceso,
2. las funciones fijas $f, g^s, h^{r,s} \in \mathbb{H}$ representan, respectivamente, el efecto del primer y segundo factor, y la interacción entre ambos.
3. Las $Y_i^{r,s} \in \mathbb{R}$ son cantidades aleatorias y conocidas que influyen el proceso según el peso de la función fija y conocida $\gamma \in \mathbb{H}$.
4. Las trayectorias aleatorias $\epsilon_i^{r,s}(t) \in \mathbb{H}$ se asumen independientes y centradas. Además, para cada par (r, s) fijo, $\epsilon_i^{r,s} \sim \text{iid } \forall i \in \{1, \dots, n_{r,s}\}$

Descripción del test

Las hipótesis de interés son:

$$H_0^A : f^1 = \dots = f^R = 0 \quad (\text{el primer factor no tiene efecto})$$

$$H_0^B : g^1 = \dots = g^S = 0 \quad (\text{el segundo factor no tiene efecto})$$

$$H_0^I : h^{1,1} = \dots = h^{R,S} = 0 \quad (\text{no hay interacción})$$

$$H_0^C : \gamma = 0 \quad (\text{la covariable no es significativa})$$

Consideraremos especialmente H_0^A . Por CFR, si existen $r_1, r_2 : f^{r_1} \neq f^{r_2}$, para cualquier medida μ gaussiana no-degenerada en \mathbb{H} ,

$$\mu \{ v \in \mathbb{H} : \langle v, f^1 \rangle = \dots = \langle v, f^R \rangle \} = 0$$

Sea v un elemento elegido al azar según μ . Luego, condicional a que H_0^A se cumple, para cada $v \in \mathbb{H}$

$$H_0^{A,v} : \langle v, f^1 \rangle = \dots = \langle v, f^R \rangle = 0$$

también se cumple, y si H_0^A no se cumple, $\mu - \text{c.s.}$ $H_0^{A,v}$ tampoco.

Consideraciones prácticas

Supuestos del modelo

- Al igual que en ANOVA clásico, las hipótesis de homocedasticidad y/o gaussianidad de los datos son cruciales. Para elegir qué test ANOVA aplicar, podemos analizar las proyecciones aleatorias.
- En funciones aleatorias, la homocedasticidad no es supuesto razonable, ya que las oscilaciones del proceso suelen depender de sus valores, así que conviene tener a mano un test ANOVA unidimensional que funcione bien bajo condiciones de heterocedasticidad.

Potencia y estabilidad

- El reemplazo de una función $\in \mathbb{H}$ por un único número real acarrea pérdida de información, y por ende de potencia para detectar alternativas
- Al estar basado en una proyección elegida *al azar*, de repetir el procedimiento dos o más veces podemos obtener resultados diferentes.

False Discovery Rate

Para reducir estos inconvenientes, se pueden tomar $k > 1$ proyecciones aleatorias, testear H_0 bajo c.u. y combinar los p -valores obtenidos de alguna forma: bootstrap (muy lenta), Bonferroni (muy conservadora) o *false discovery rate* (FDR).

Si se testean k diferentes hipótesis, La “*tasa de falso descubrimiento*” es la proporción esperada de hipótesis incorrectamente rechazadas. Si se testea k veces la misma hipótesis, la FDR coincide con el nivel de significación del test.

En particular, si $p_{(1)} \leq \dots \leq p_{(k)}$ son los k p -valores (ordenados), el nivel de un test que rechace H_0 cuando $\alpha \geq \inf \left\{ \frac{k}{i} p_{(i)}, i = 1, \dots, k \right\}$ es, a lo sumo, de α .

FDR vs. Bonferroni

Bonferroni rechaza con nivel α cuando $p_{(1)} \leq \alpha/k$, en cuyo caso FDR también rechaza. En la práctica, FDR suele ser mucho menos conservador que Bonferroni.

Consideraciones prácticas: ¿cuántas direcciones tomar?

Resta definir el número de proyecciones. $k = 30$ es más que suficientemente conservador, y en algunos casos bastará con $k \in \{1, \dots, 5\}$. Es importante considerar que a mayor k ,

- Bajo H_0 , más conservador resulta el test y menor será la probabilidad de rechazo
- bajo una alternativa fija H_1 , mayor será la probabilidad de rechazo

Este último efecto puede deberse a que mientras más direcciones aleatorias se tomen, más altas son las chances de que la alternativa correlacione fuertemente con alguna de ellas y este efecto contrarreste lo conservador del procedimiento. Por ello, en [Cuesta 2010] recomiendan tomar $k = \min \{30, n\}$.

En R: usando fda.usc::fanova.RPm









```
library(fda.usc)
```

```
X <- cbind(fda::growth$hgtm, fda::growth$hgtf)
grilla <- fda::growth$age
factores <- data.frame(
  varon=as.factor(startsWith(colnames(X) , "boy"))
)
```

```
X.fd <- fdata(mdata=t(X), argvals=grilla)
test <- fanova.RPm(X.fd, ~ varon, factores, RP=1)
test$p.FDR
```

```
##                varon
## RP1 2.038658e-11
```

Referencias

-  Cuesta-Albertos, J. A., del Barrio, E., Fraiman, R., & Matrán, C. (2007). The random projection method in goodness of fit for functional data. *Computational Statistics & Data Analysis*, 51(10), 4814-4831.
-  Cuevas, A., Febrero, M., & Fraiman, R. (2007). Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 22(3), 481-496.
-  Cuesta-Albertos, J. A., & Febrero-Bande, M. (2010). A simple multiway ANOVA for functional data. *Test*, 19(3), 537-557.
-  Patilea, V., Sanchez-Sellero, C., & Saumard, M. (2012). Projection-based nonparametric goodness-of-fit testing with functional covariates. *arXiv preprint arXiv:1205.5578*.
-  García-Portugués, E., González-Manteiga, W., & Febrero-Bande, M. (2014). A goodness-of-fit test for the functional linear model with scalar response. *Journal of Computational and Graphical Statistics*, 23(3), 761-778.
-  Cuesta-Albertos, J. A., Febrero-Bande, M., & de la Fuente, M. O. (2017). The DD^G -classifier in the functional setting. *Test*, 26(1), 119-142.
-  Cuesta-Albertos, J. A., García-Portugués, E., Febrero-Bande, M., & González-Manteiga, W. (2019). Goodness-of-fit tests for the functional linear model based on randomly projected empirical processes. *The Annals of Statistics*, 47(1), 439-467.
-  Librería fda.usc

¡Gracias!

github.com/memfcen-amateur/proyecciones-aleatorias