



Functional data analysis: interpolation, registration, and nearest neighbors in *scikit-fda*

Pablo Marcos Manchón





Análisis de datos funcionales: interpolación, registro y vecinos próximos en *scikit-fda*

Pablo Marcos Manchón





Análisis de datos funcionales (FDA)

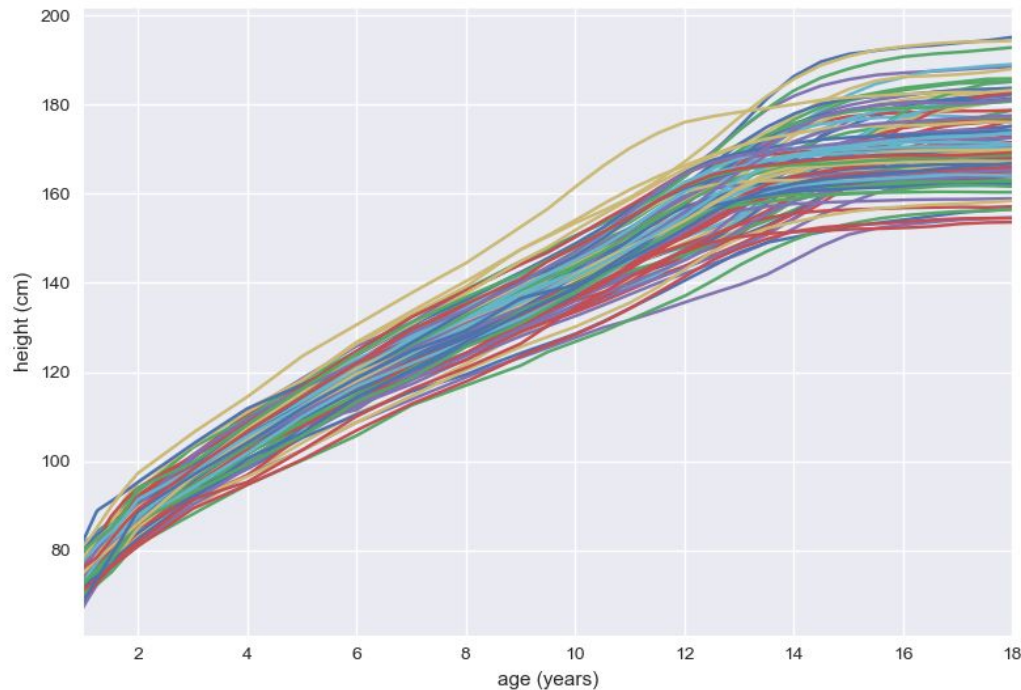
- Estudio de datos de naturaleza continua
- Los datos forman un conjunto de funciones $\{f_i\}_{i=1}^n$
- Aplicaciones en medicina, bioinformática, ingeniería...

Datos funcionales

- Datos univariantes

$$f_i : \mathbb{R} \rightarrow \mathbb{R}$$

- Curvas
- Superficies
- Datos multivariantes



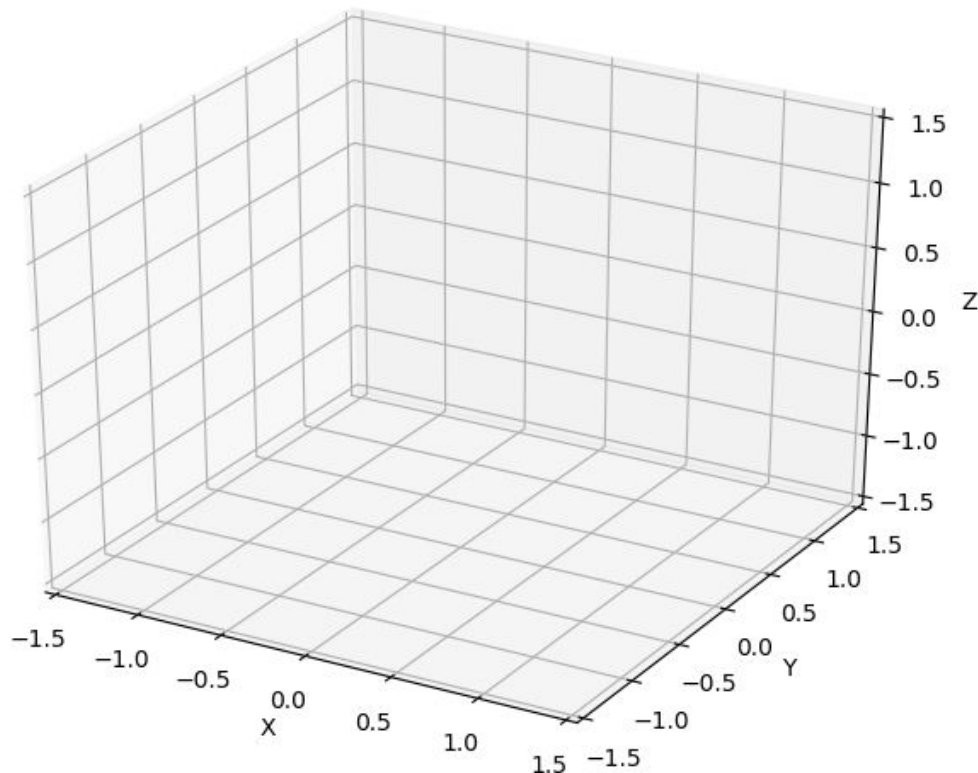
Datos funcionales

- Datos univariantes

- **Curvas**

$$f_i : \mathbb{R} \rightarrow \mathbb{R}^m$$

- Superficies
- Datos multivariantes



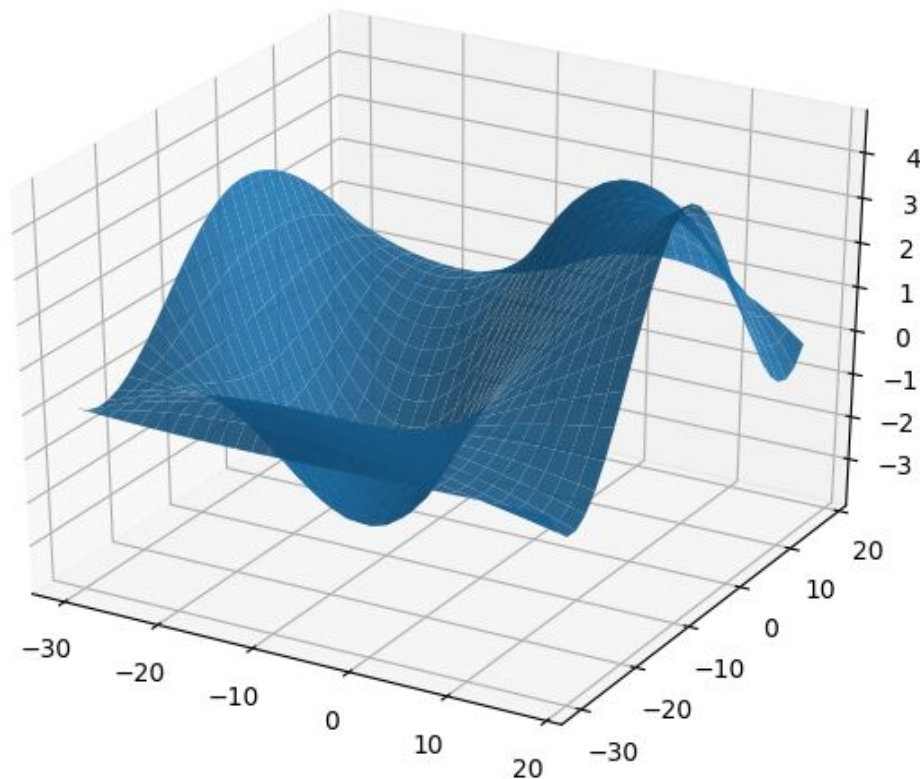
Datos funcionales

- Datos univariantes
- Curvas

- **Superficies**

$$f_i : \mathbb{R}^2 \rightarrow \mathbb{R}$$

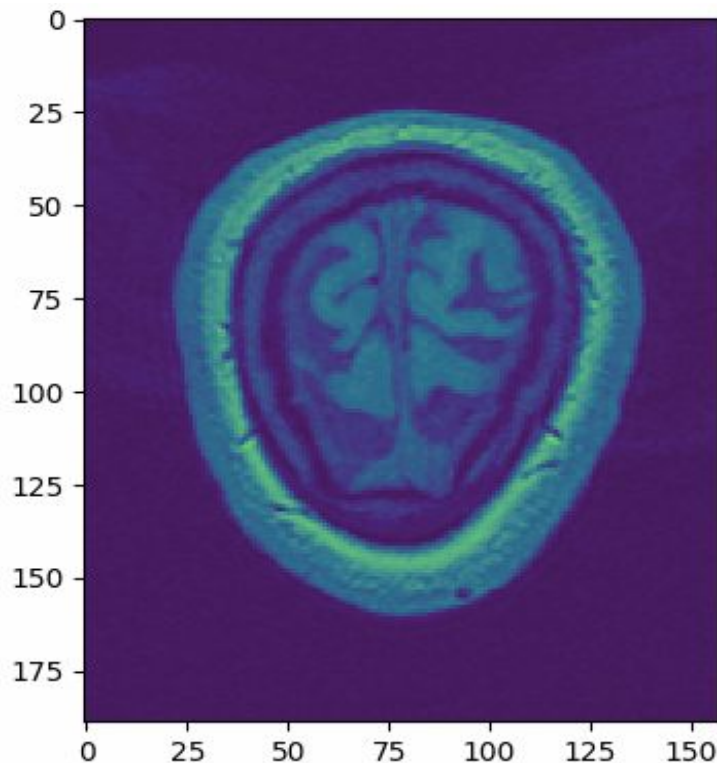
- Datos multivariantes



Datos funcionales

- Datos univariantes
- Curvas
- Superficies
- Datos multivariantes

$$f_i : \mathbb{R}^d \rightarrow \mathbb{R}^m$$



Proyecto scikit-fda

- Soporte al análisis de datos funcionales en **Python**
- Proyecto de código abierto
- Ecosistema de paquetes de computación científica de **SciPy**



scikit-fda
functional data analysis in python

Objetivo

- Extender las funcionalidades del paquete
 - Interpolación
 - Registro
 - Estimadores de vecinos próximos



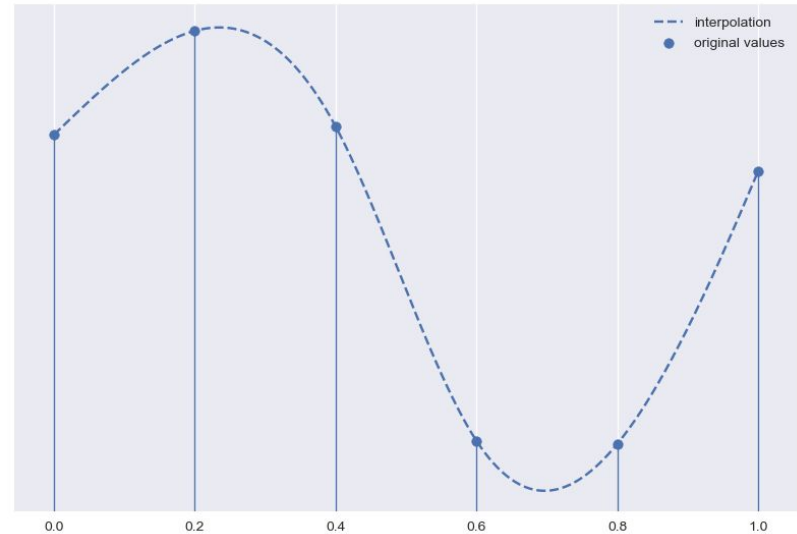
scikit-fda
functional data analysis in python

Representación	Exploración	Preprocesamiento	Inferencia	Machine learning
Paramétrica <ul style="list-style-type: none"> Bases No paramétrica <ul style="list-style-type: none"> Densa Dispersa 	Visualización Estadísticos Profundidad Datos atípicos Reducción de dimensionalidad Distancias	Suavizado Registro Derivación Transformaciones	Tests Intervalos de confianza	Clasificación Regresión Clustering

Representación	Exploración	Preprocesamiento	Inferencia	Machine learning
Paramétrica <ul style="list-style-type: none"> Bases No paramétrica <ul style="list-style-type: none"> Densa Dispersa 	Visualización Estadísticos Profundidad Datos atípicos Reducción de dimensionalidad Distancias	Suavizado Registro Derivación Transformaciones	Tests Intervalos de confianza	Clasificación Regresión Clustering

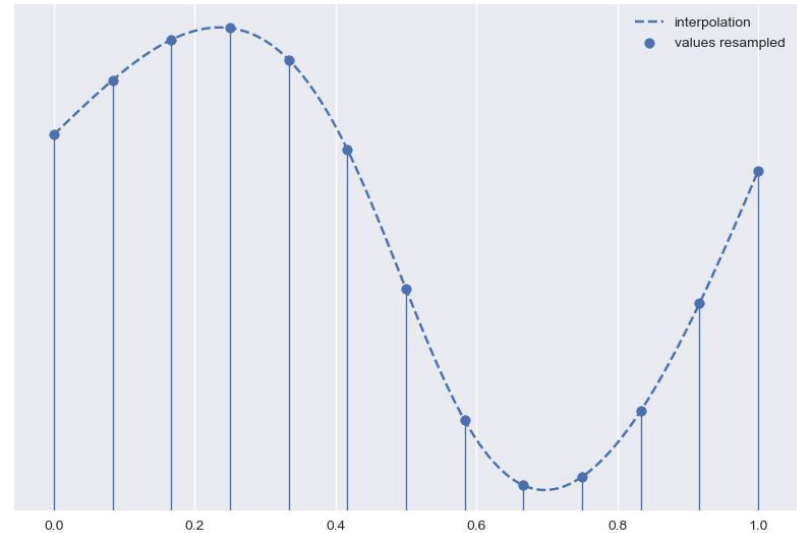
Interpolación

- Los datos son observados en valores discretos $f_i(t_j)$
- Evaluación en puntos distintos a los observados
- Métodos basado en splines



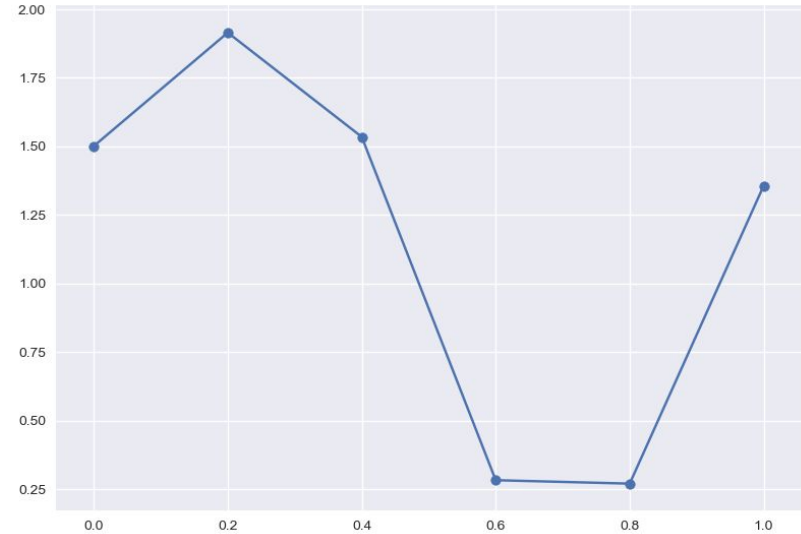
Interpolación

- Los datos son observados en valores discretos $f_i(t_j)$
- Evaluación en puntos distintos a los observados
- Métodos basado en splines



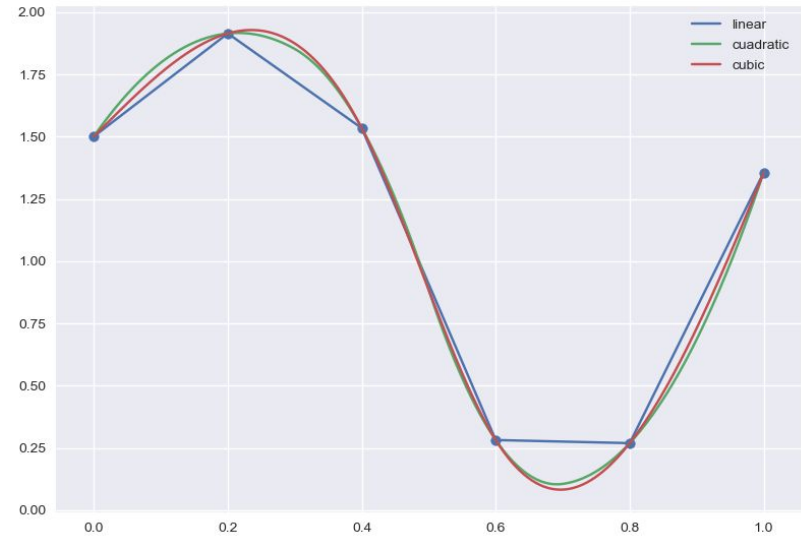
Interpolación lineal

- Uso de segmentos de línea para unir los puntos
- Función continua lineal a trozos
- Sencillo y eficiente



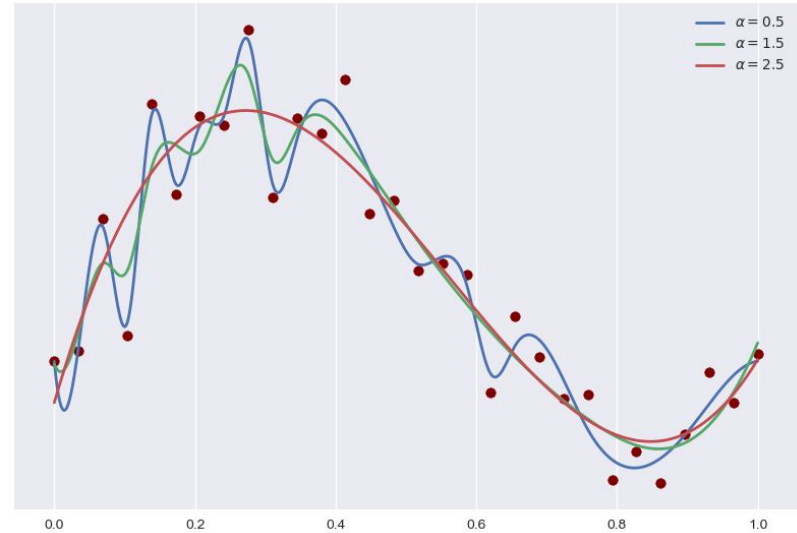
Interpolación de splines

- Uso de polinomios para unir los puntos
- Funciones definidas a trozos
- Permiten el uso de derivadas

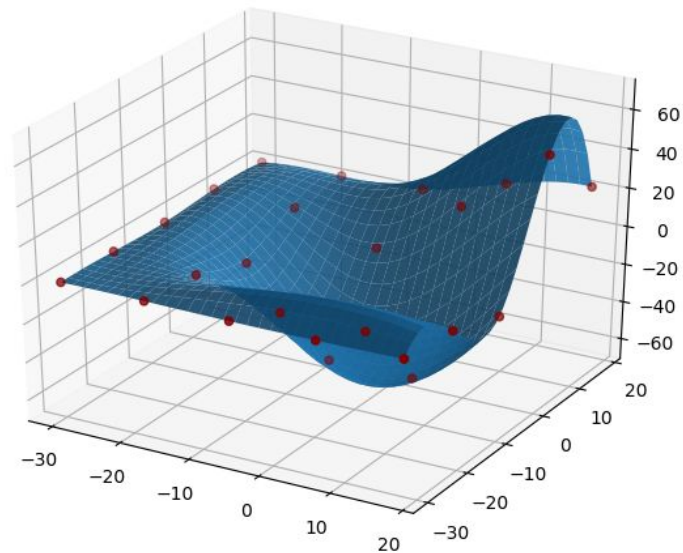
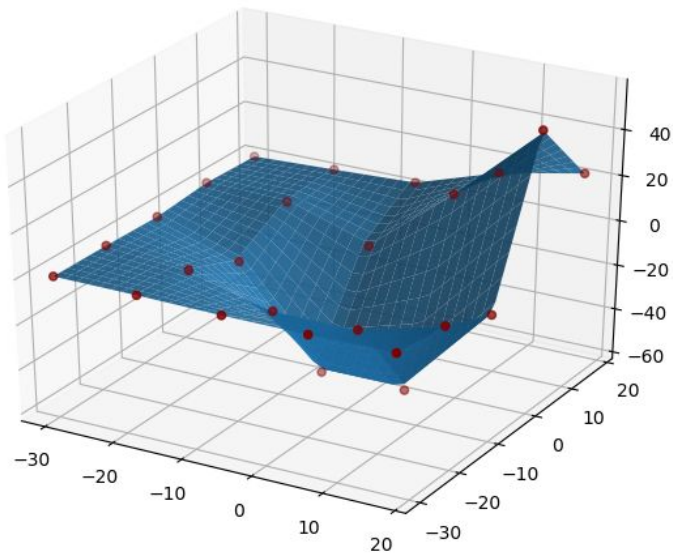


Interpolación suavizada

- Variación de la interpolación de splines
- Suavidad en las funciones
- Eliminación de ruido en mediciones



Interpolación multivariante

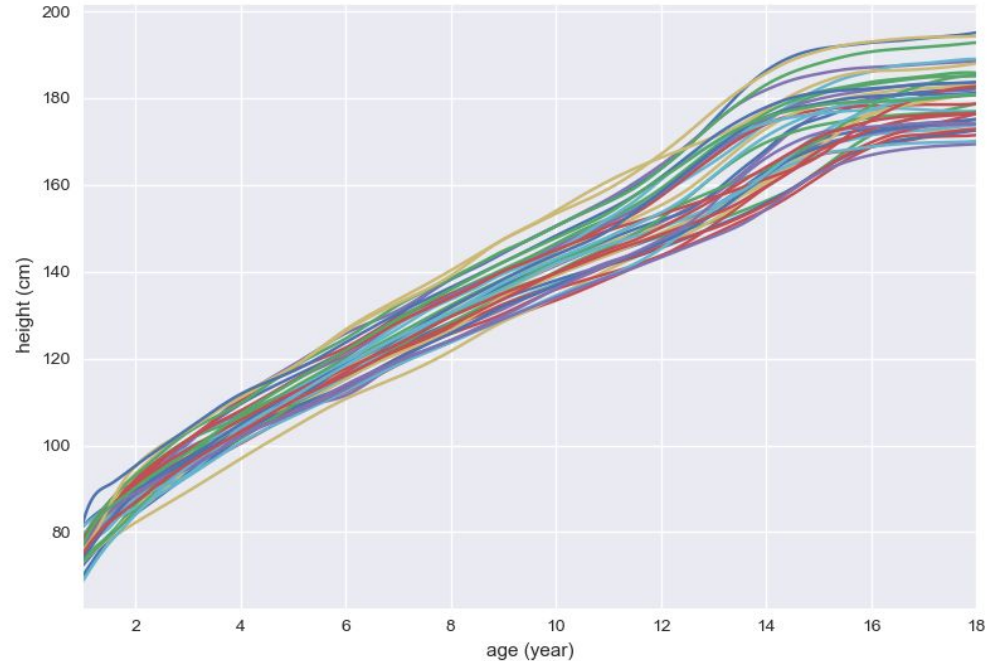


Representación	Exploración	Preprocesamiento	Inferencia	Machine learning
Paramétrica <ul style="list-style-type: none"> Bases No paramétrica <ul style="list-style-type: none"> Densa Dispersa 	Visualización Estadísticos Profundidad Datos atípicos Reducción de dimensionalidad Distancias	Suavizado Registro Derivación Transformaciones	Tests Intervalos de confianza	Clasificación Regresión Clustering

Representación	Exploración	Preprocesamiento	Inferencia	Machine learning
Paramétrica <ul style="list-style-type: none"> Bases No paramétrica <ul style="list-style-type: none"> Densa Dispersa 	Visualización Estadísticos Profundidad Datos atípicos Reducción de dimensionalidad Distancias	Suavizado Registro Derivación Transformaciones	Tests Intervalos de confianza	Clasificación Regresión Clustering

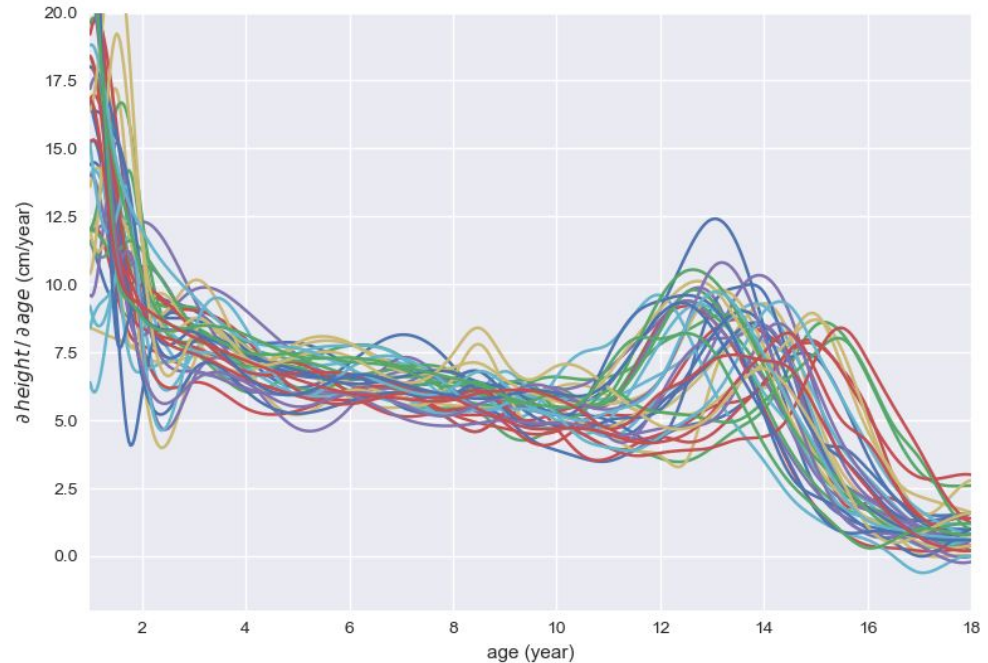
Registro

- Muestras presentan formas similares, pero no alineadas
- La variabilidad procede de su dominio
- Es deseable cuantificar y reducir esta variación



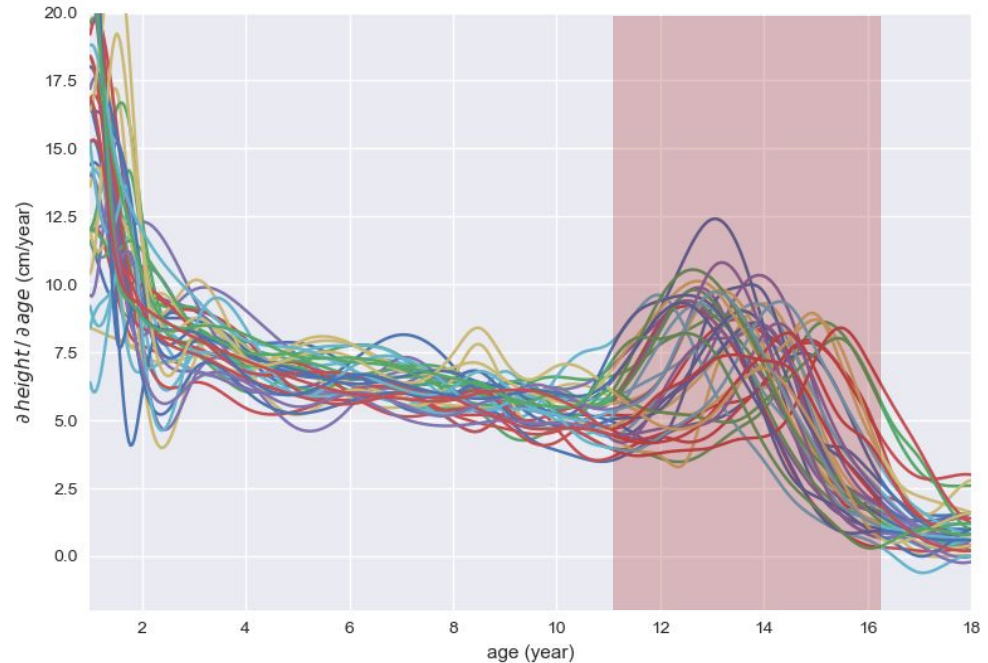
Registro

- Muestras presentan formas similares, pero no alineadas
- La variabilidad procede de su dominio
- Es deseable cuantificar y reducir esta variación

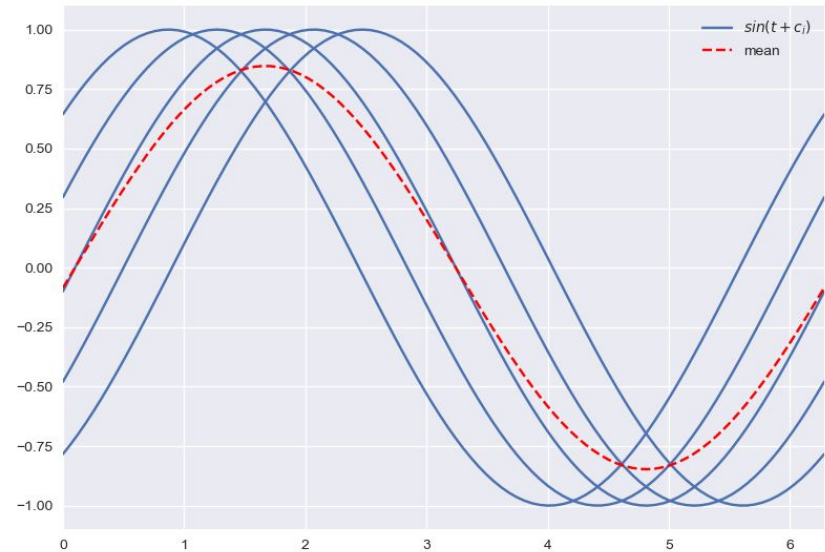
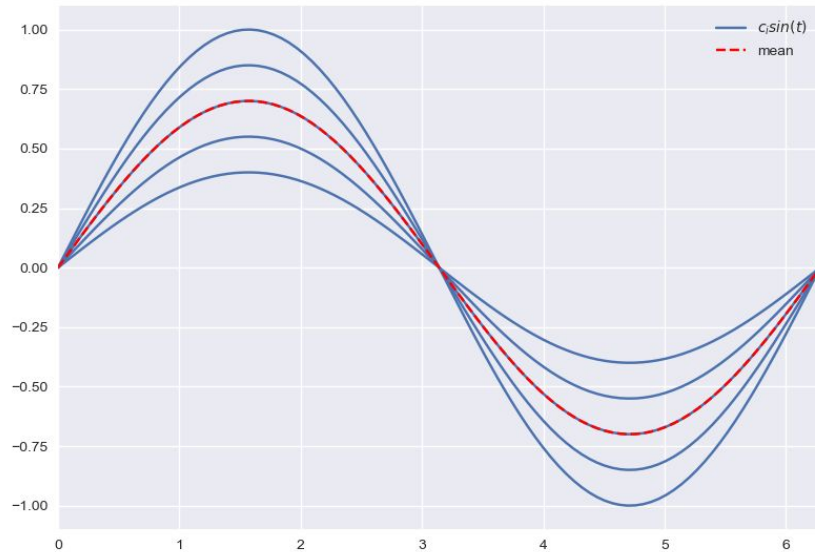


Registro

- Muestras presentan formas similares, pero no alineadas
- La variabilidad procede de su dominio
- Es deseable cuantificar y reducir esta variación



Amplitud y fase



Traslaciones

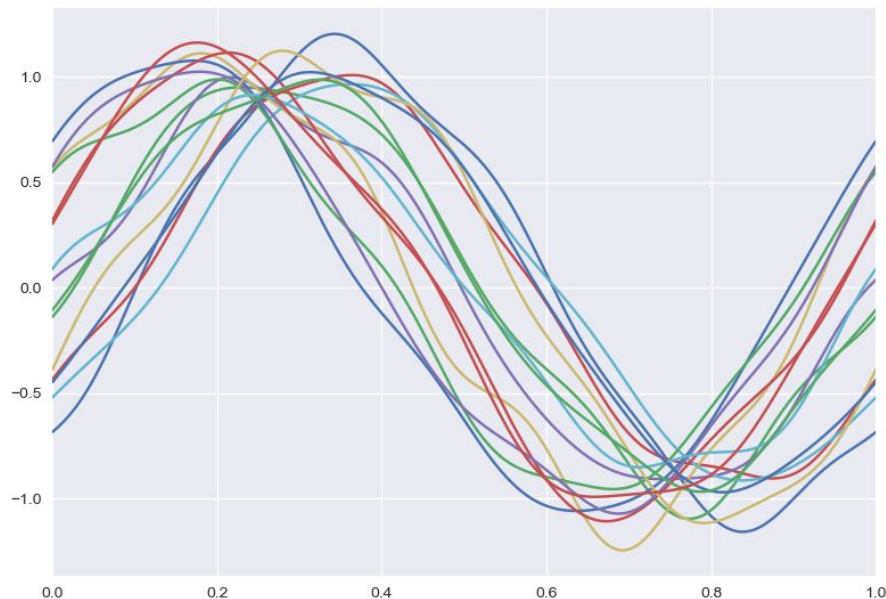
- Denominado *shift registration*

- Uso de traslaciones

$$f_i^*(t) = f_i(t + \delta_i)$$

- Minimización de suma de errores cuadrados *REGSSE*

$$\int \sum_{i=1}^n (f_i^*(t) - \bar{f}^*(t))^2 dt$$



Traslaciones

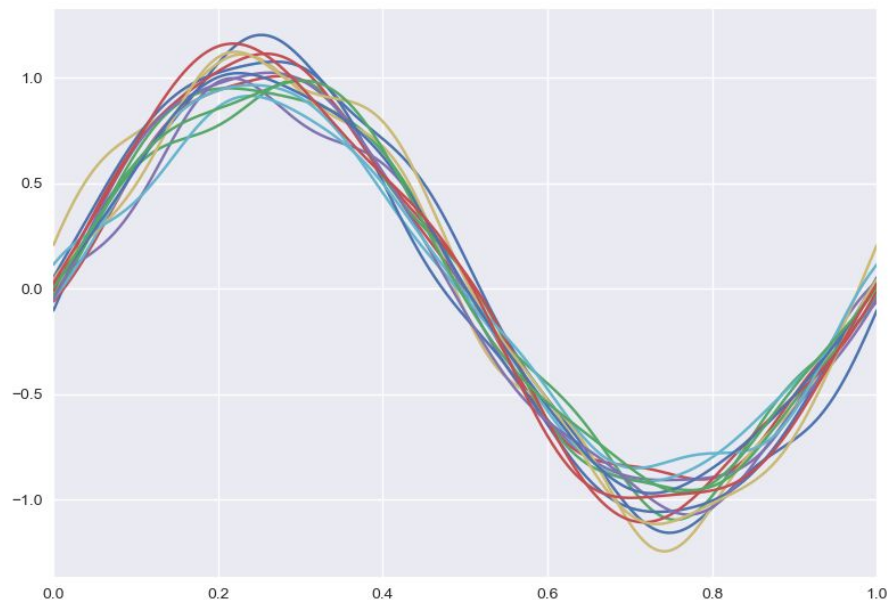
- Denominado *shift registration*

- Uso de traslaciones

$$f_i^*(t) = f_i(t + \delta_i)$$

- Minimización de suma de errores cuadrados *REGSSE*

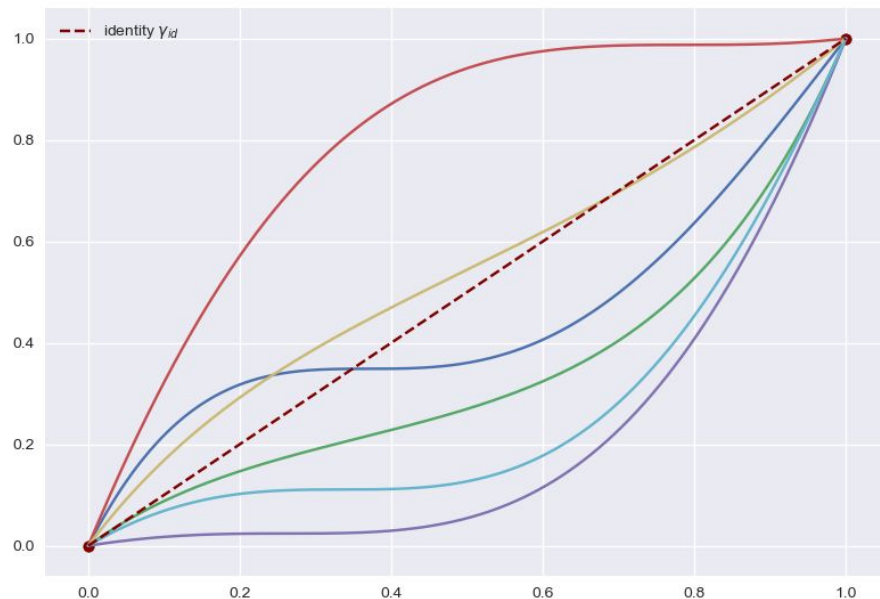
$$\int \sum_{i=1}^n (f_i^*(t) - \bar{f}^*(t))^2 dt$$

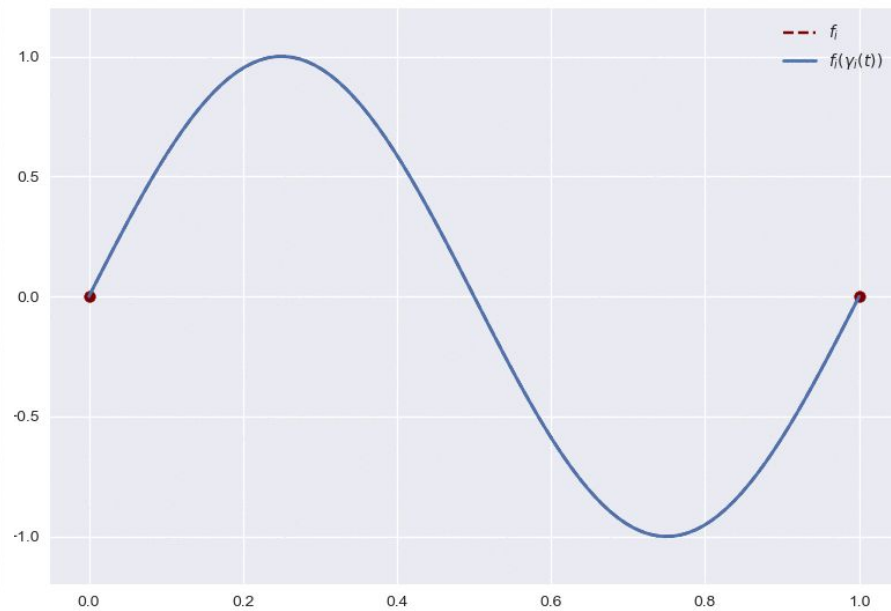
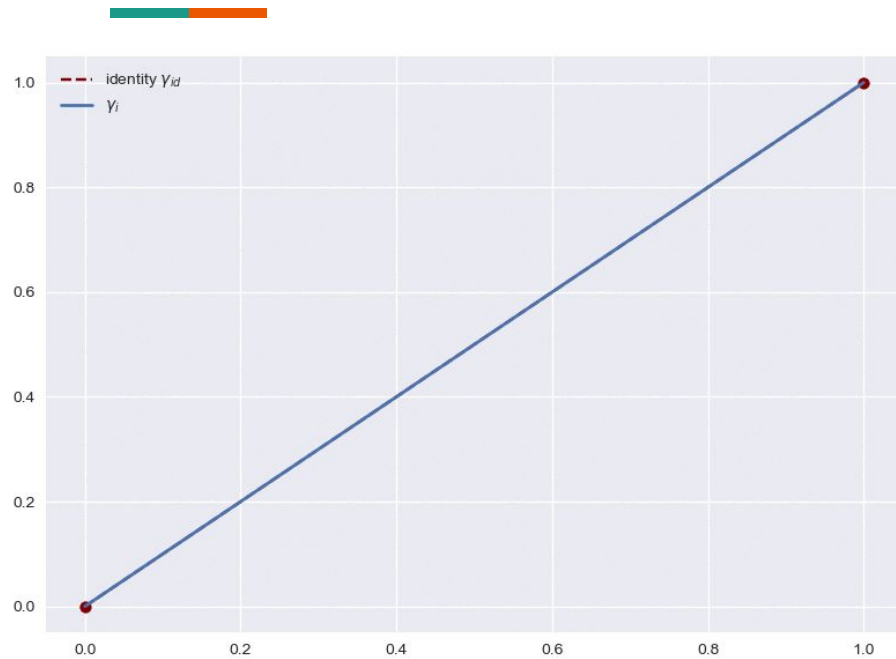


Transformaciones generales

- Difeomorfismos del dominio
- Representados por *warpings*
 - $\gamma_i(t) : \mathcal{T} \rightarrow \mathcal{T}$
 - Continuas y crecientes
 - Frontera invariante
- Composición de funciones

$$f_i^*(t) = f_i(\gamma_i(t)) = f_i \circ \gamma_i$$

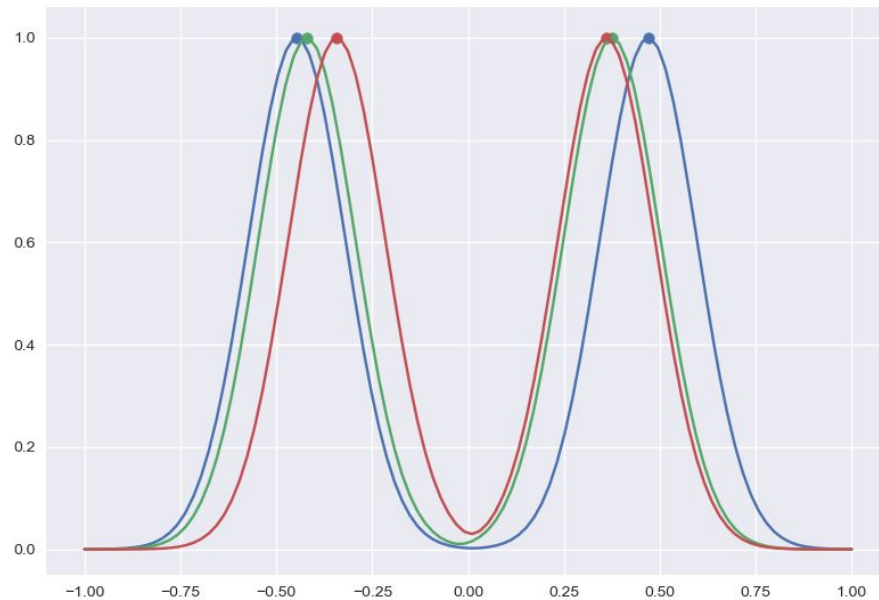




Puntos de referencia

- Denominado *landmark registration*
- Puntos característicos (e. g. máximos o mínimos)
- Alineación a puntos de referencia

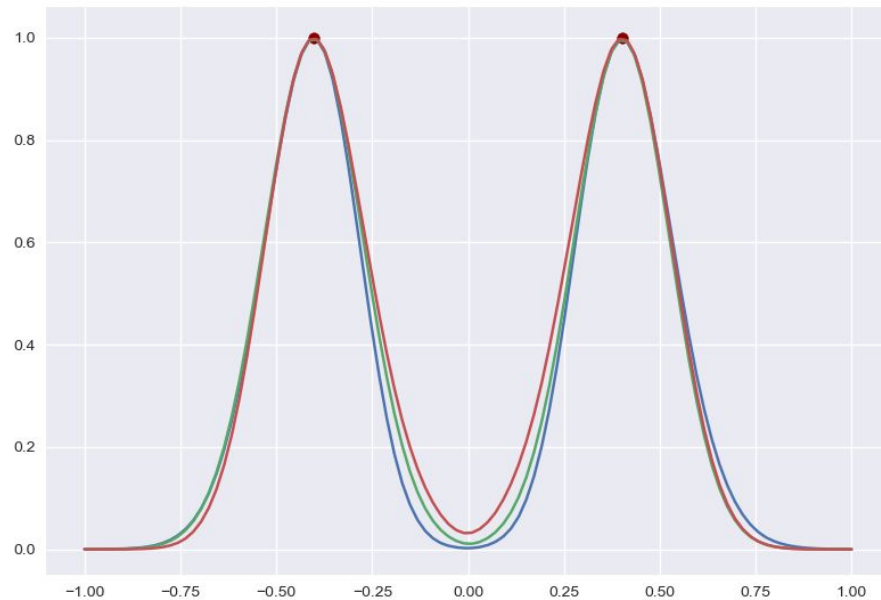
$$\gamma_i(t_j^*) = t_{ij}$$



Puntos de referencia

- Denominado *landmark registration*
- Puntos característicos (e. g. máximos o mínimos)
- Alineación a puntos de referencia

$$\gamma_i(t_j^*) = t_{ij}$$

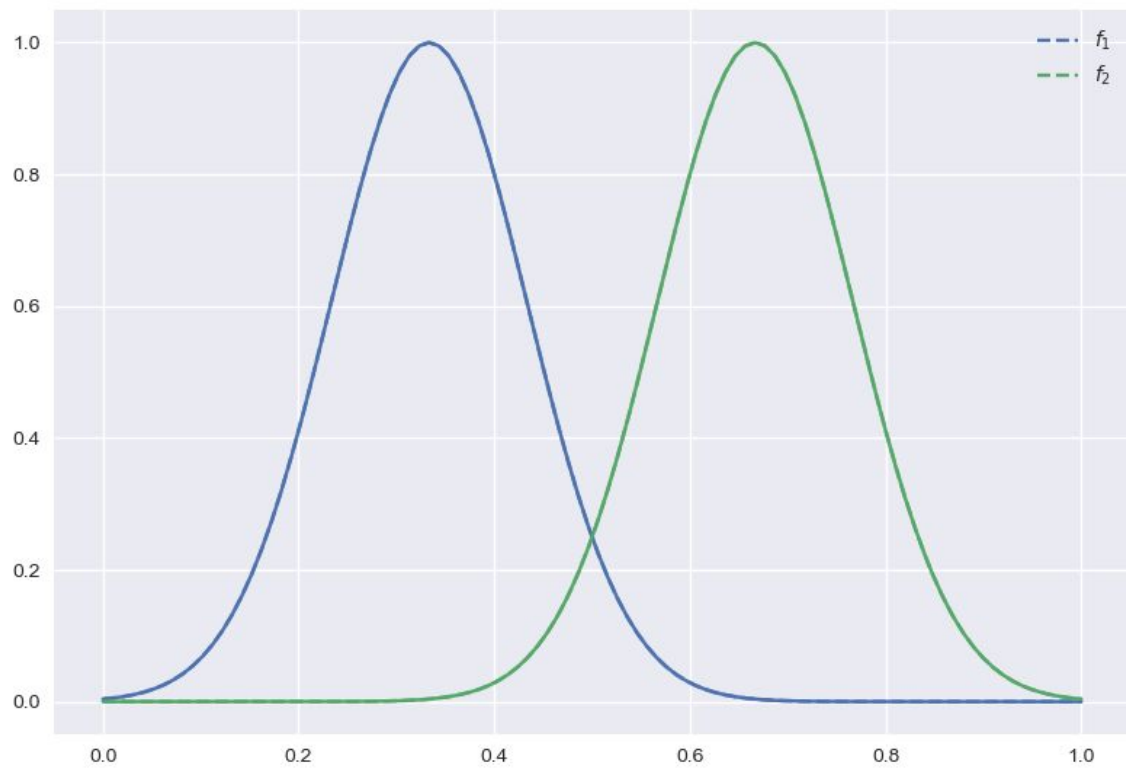


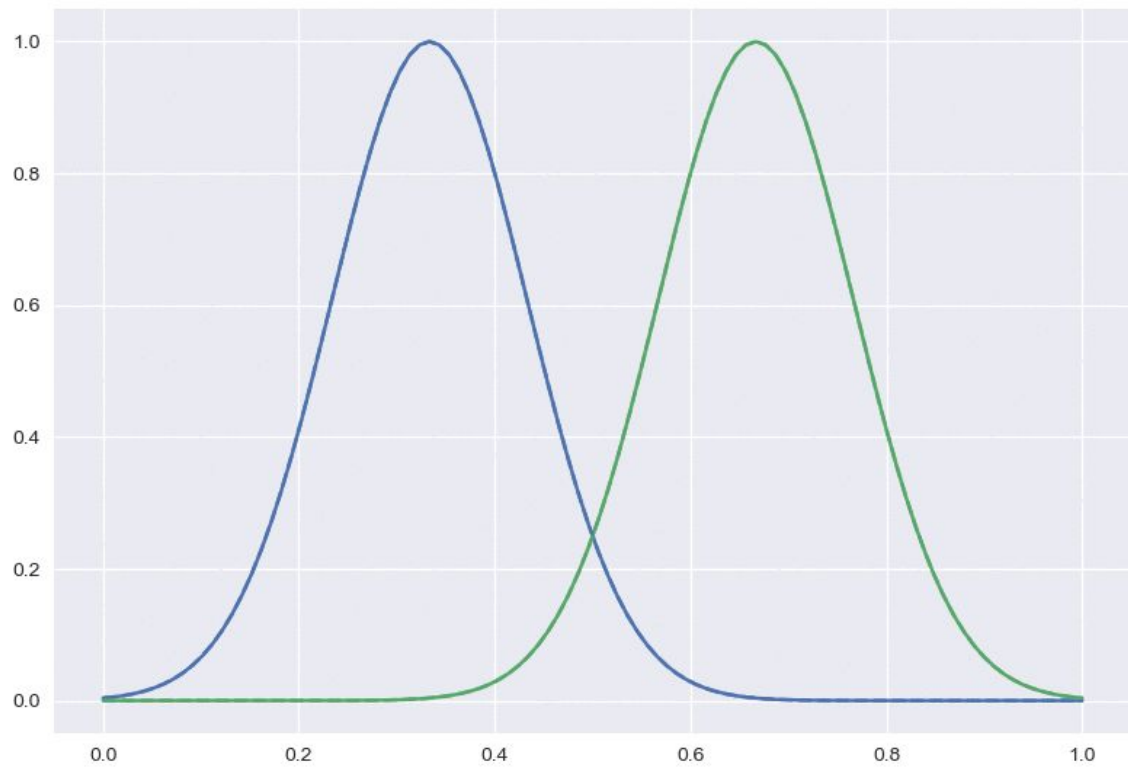


Alineación por pares

- Búsqueda criterio empleando la estructura continua
- No supervisado
- Minimización de un funcional de energía para la alineación de dos muestras

$$\gamma_{21} = \operatorname{argmín}_{\gamma \in \Gamma} E[f_1, f_2 \circ \gamma]$$







Registro de un conjunto de muestras

- Creación de una plantilla $\mu(t)$
- Todas las muestras son alineadas a esta plantilla común

$$\gamma_i = \underset{\gamma \in \Gamma}{\operatorname{argmín}} E[\mu, f_i \circ \gamma]$$

- Término de energía y plantilla adecuados



Registro elástico

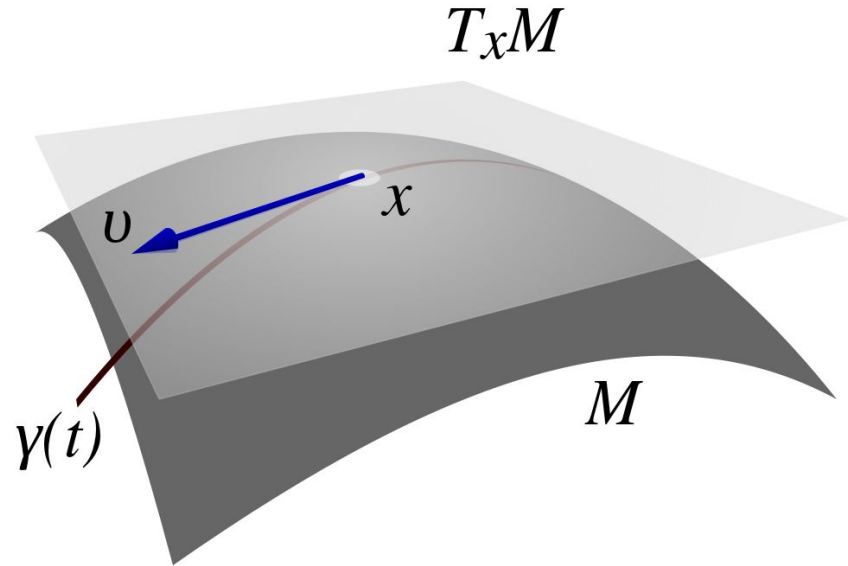
- Empleo de geometría de Riemann
- Métrica de Fisher-Rao como energía
- Invariancia respecto a deformaciones simultáneas

$$d_{FR}(f_1, f_2) = d_{FR}(f_1 \circ \gamma, f_2 \circ \gamma)$$

- Marco matemático denominado análisis elástico

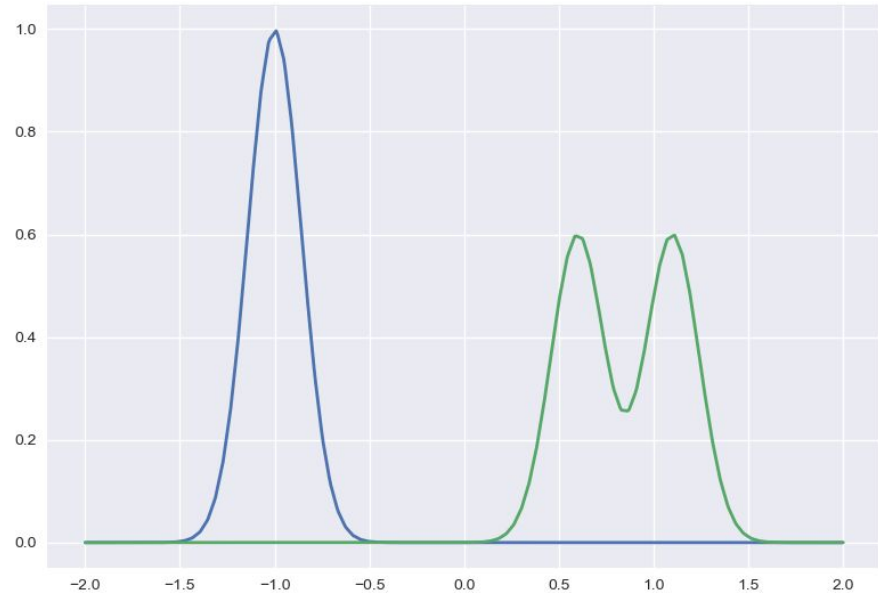
Variedad de Riemann

- Métrica de Riemann
- Definida en el espacio tangente
- Longitud de camino geodésico



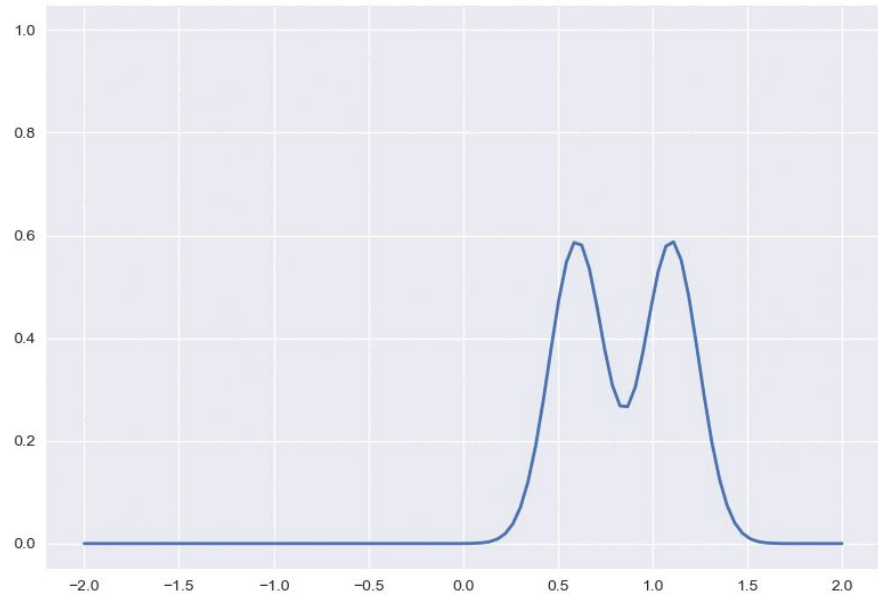
Variedad de Riemann

- Métrica de Riemann
- Definida en el espacio tangente
- Longitud de camino geodésico



Variedad de Riemann

- Métrica de Riemann
- Definida en el espacio tangente
- Longitud de camino geodésico



Transformada SRSF

- La transformada SRSF aplanar la variedad

$$SRSF\{f\} = \text{sign}(\dot{f})\sqrt{|\dot{f}|}$$

- Convierte la métrica de Fisher-Rao en la métrica \mathbb{L}^2

$$\|q_1 - q_2\|_{\mathbb{L}^2} = \int |q_1(t) - q_2(t)|^2 dt$$

- Permite computación eficiente

$$\begin{array}{ccc} f & \xrightarrow{\text{SRSF}} & q \\ \text{action on } \mathcal{F} \downarrow & & \downarrow \text{action on } \mathbb{L}^2 \\ f \circ \gamma & \xrightarrow{\text{SRSF}} & (q, \gamma) \end{array}$$

Transformada SRSF

- Transformación de funciones

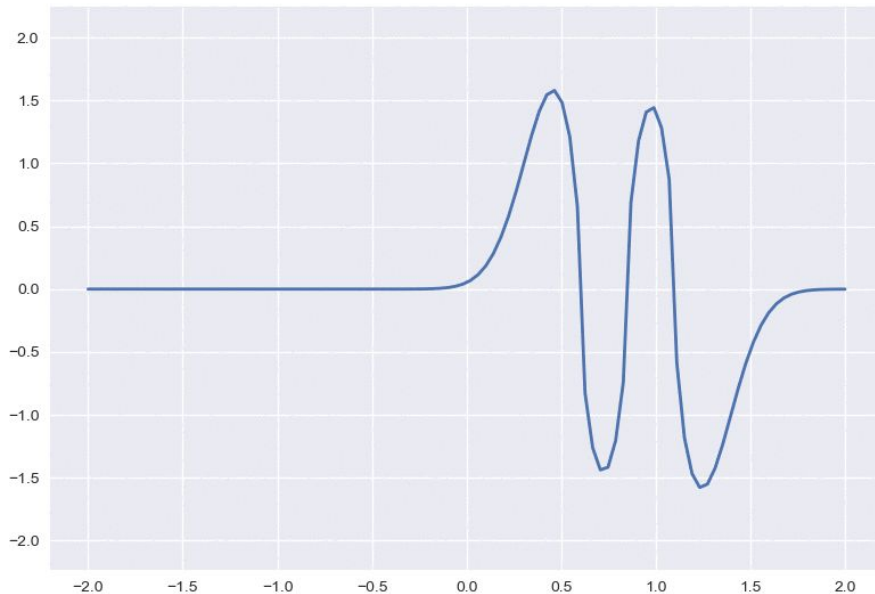
$$SRSF\{f\} = \text{sign}(\dot{f})\sqrt{|\dot{f}|}$$

- Aplana la variedad

$$\|q_1 - q_2\|_{\mathbb{L}^2}^2 = \int |q_1(t) - q_2(t)|^2 dt$$

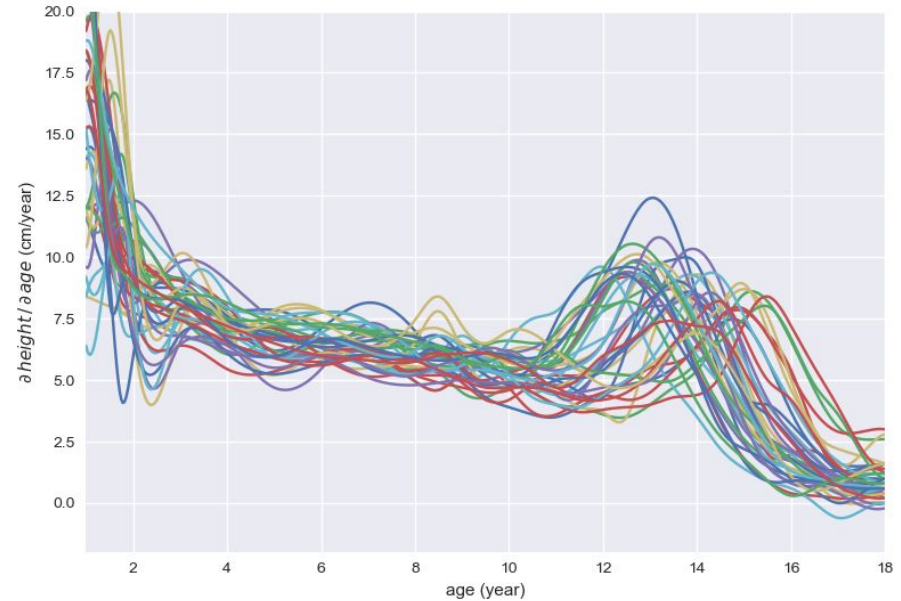
- Las geodésicas son rectas

$$L[\alpha] = \alpha q_1 + (1 - \alpha)q_2$$



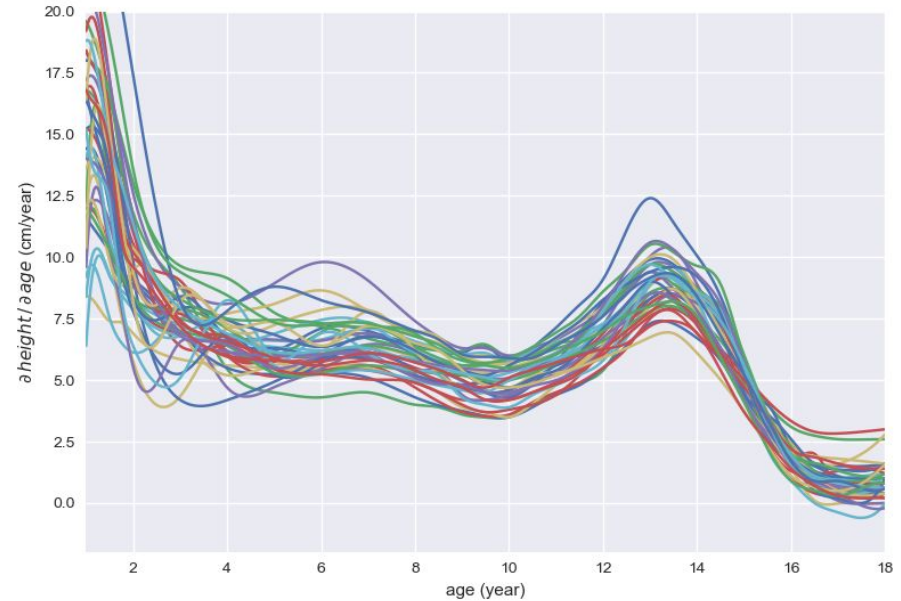
Registro elástico

- Uso de la *media elástica* como plantilla $\mu(t)$
- Métrica de Fisher-Rao como energía



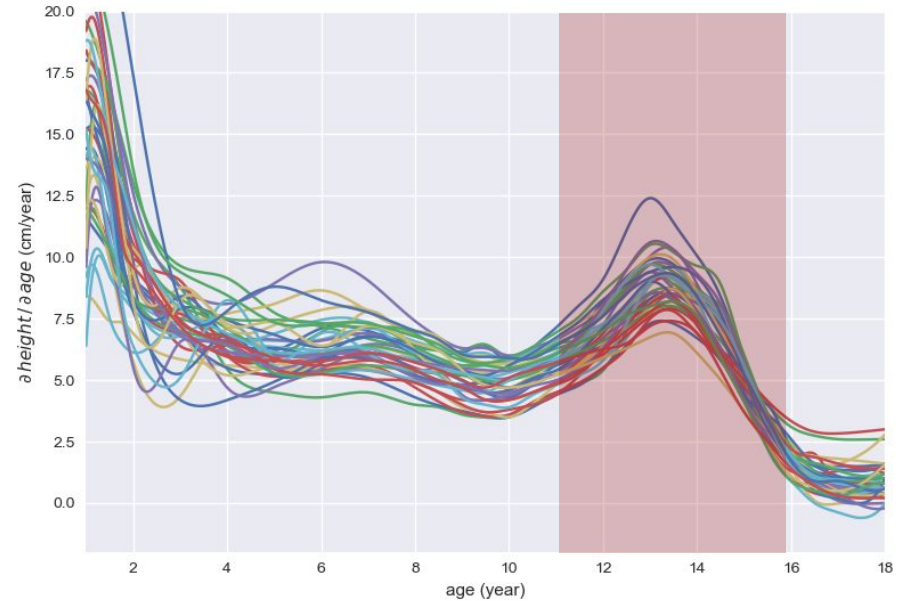
Registro elástico

- Uso de la *media elástica* como plantilla $\mu(t)$
- Métrica de Fisher-Rao como energía



Registro elástico

- Uso de la *media elástica* como plantilla $\mu(t)$
- Métrica de Fisher-Rao como energía

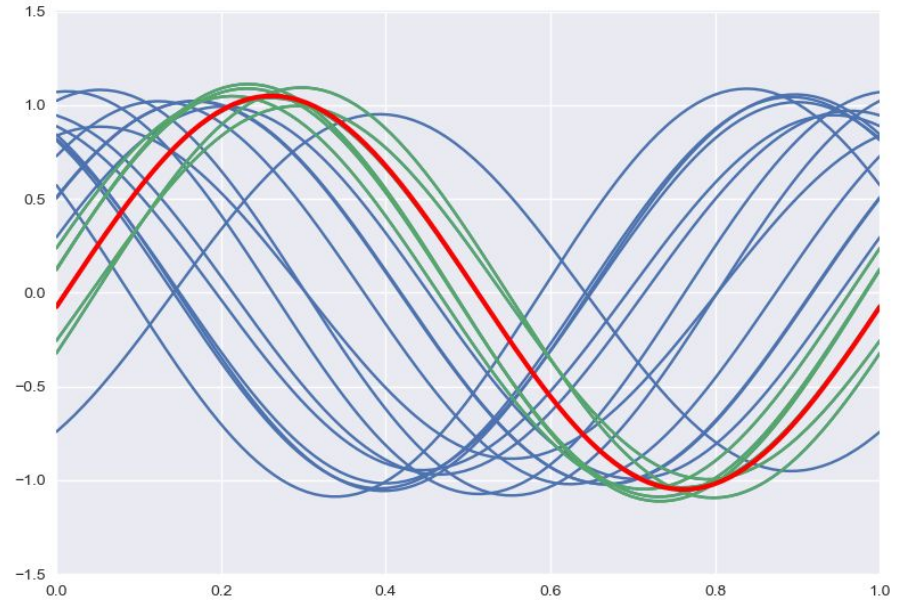


Representación	Exploración	Preprocesamiento	Inferencia	Machine learning
Paramétrica <ul style="list-style-type: none"> Bases No paramétrica <ul style="list-style-type: none"> Densa Dispersa 	Visualización Estadísticos Profundidad Datos atípicos Reducción de dimensionalidad Distancias	Suavizado Registro Derivación Transformaciones	Tests Intervalos de confianza	Clasificación Regresión Clustering

Representación	Exploración	Preprocesamiento	Inferencia	Machine learning
Paramétrica <ul style="list-style-type: none"> • Bases No paramétrica <ul style="list-style-type: none"> • Densa • Dispersa 	Visualización Estadísticos Profundidad Datos atípicos Reducción de dimensionalidad Distancias	Suavizado Registro Derivación Transformaciones	Tests Intervalos de confianza	Clasificación Regresión Clustering

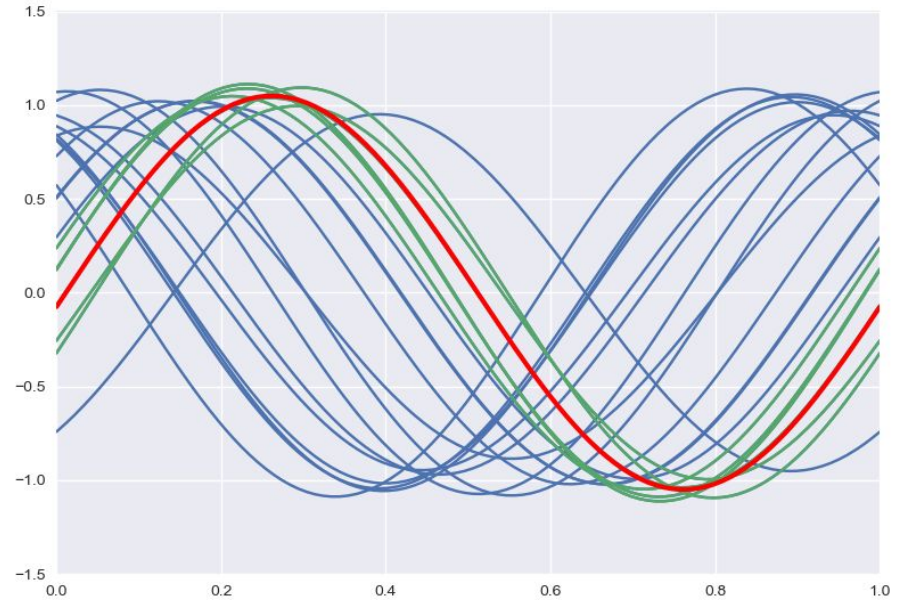
Vecinos próximos

- Generalización a espacios funcionales
- Emplean noción de localidad
- Usados en problemas de clasificación y regresión



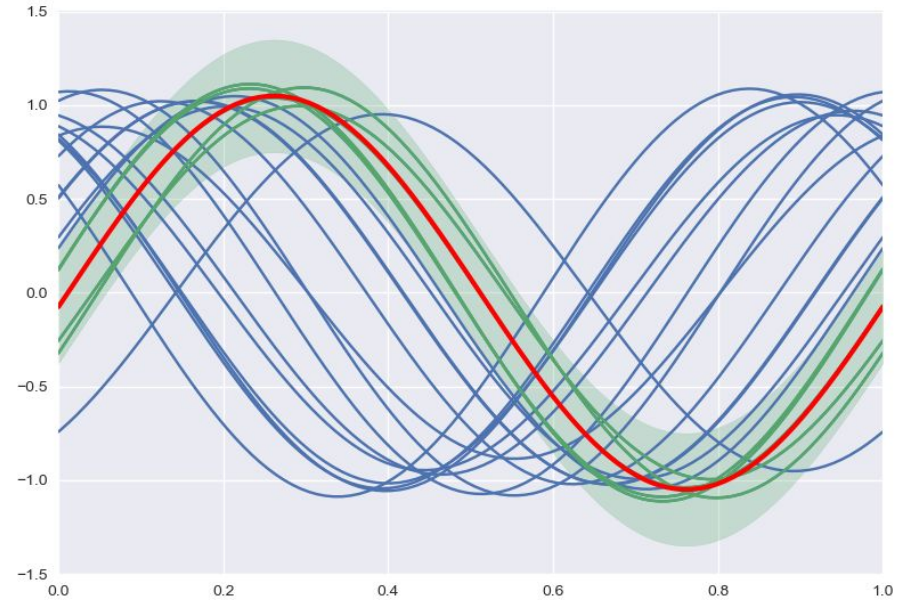
Vecinos próximos

- K-NN
 - Uso de los K vecinos más próximos
- Radius-NN
 - Muestras a menor distancia que un radio



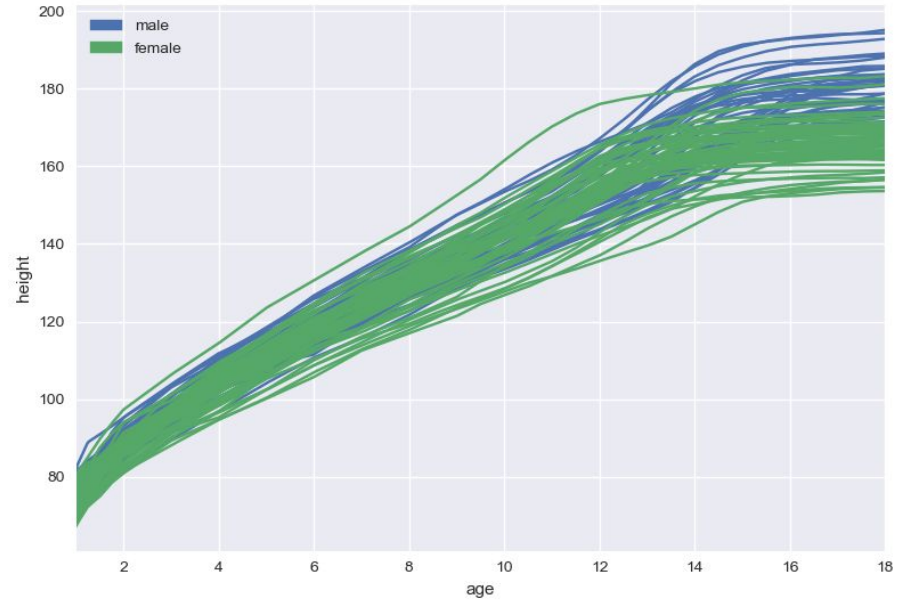
Vecinos próximos

- K-NN
 - Uso de los K vecinos más próximos
- Radius-NN
 - Muestras a menor distancia que un radio



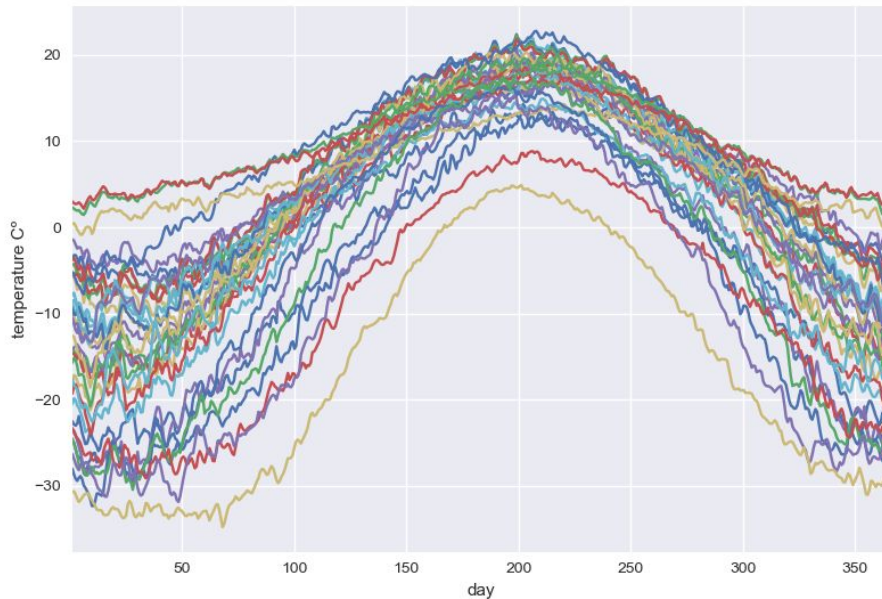
Clasificación

- Cada muestra de entrenamiento tiene una etiqueta de clase
- Voto ponderado de los vecinos



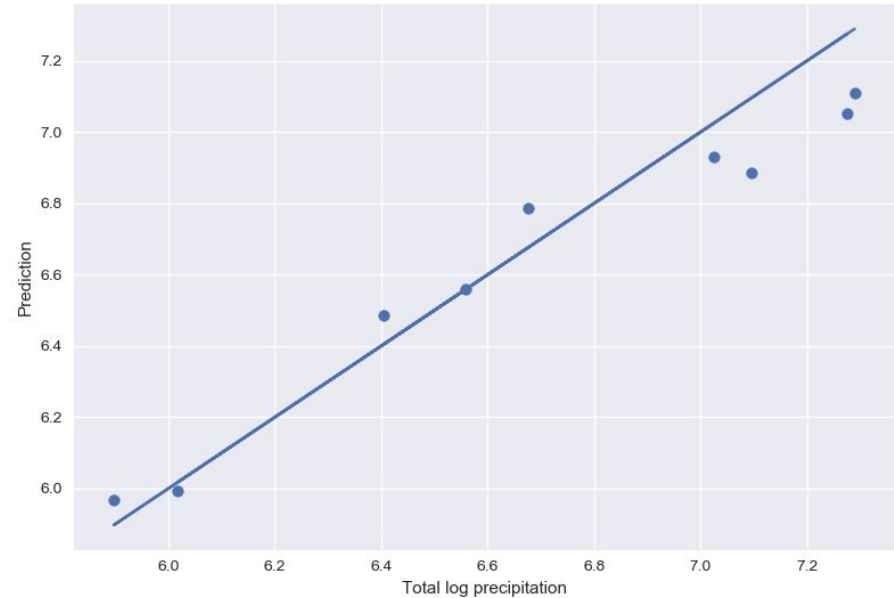
Regresión

- Cada muestra de entrenamiento tiene asociada una respuesta
- Predicción de la respuesta
 - Respuesta escalar
 - Respuesta funcional



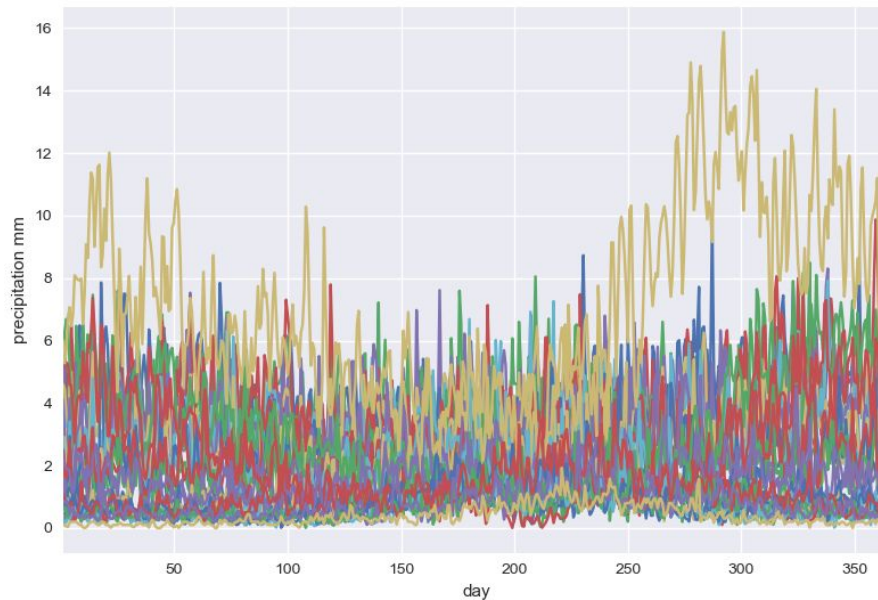
Regresión

- Cada muestra de entrenamiento tiene asociada una respuesta
- Predicción de la respuesta
 - **Respuesta escalar**
 - Respuesta funcional



Regresión

- Cada muestra de entrenamiento tiene asociada una respuesta
- Predicción de la respuesta
 - Respuesta escalar
 - **Respuesta funcional**



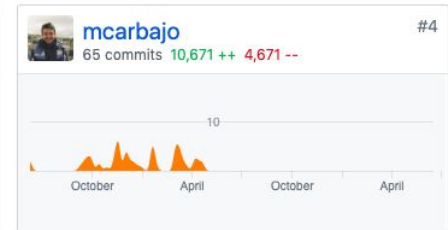
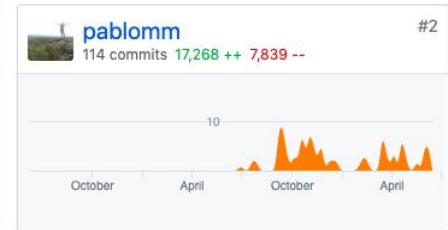
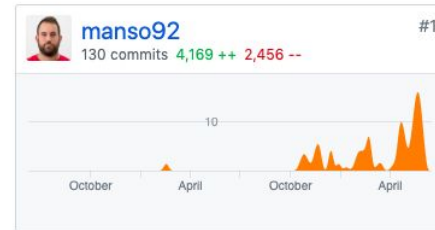
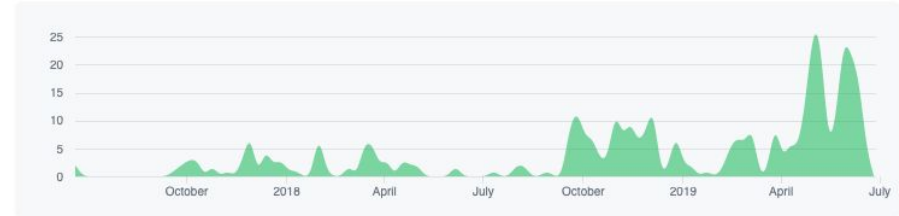
Desarrollo

- Código público en *Github*
- Estándares de codificación
- Trabajo en equipo
- Documentación
- Pruebas
- Integración continua



Desarrollo

- Código público en *Github*
- Estándares de codificación
- Trabajo en equipo
- Documentación
- Pruebas
- Integración continua



Desarrollo

- Código público en *Github*
- Estándares de codificación
- Trabajo en equipo
- Documentación
- Pruebas
- Integración continua

Operations with image dimensions #101

 Merged pablomm merged 16 commits into [develop](#) from [feature/image-operations](#) 15 days ago

 Conversation 40

 Commits 16

 Checks 0

 Files changed 4

+393 -41



pablomm commented on 17 May

Member



- Method `image` to extract an image dimension. Examples of the functionality.

Get a component of the samples (f1, f2, f3)

```
>>> fd = make_multimodal_samples(n_samples=15, ndim_image=3)
>>> fd.ndim_image
3
>>> fd_1 = fd.image(1) # Extracts f1
>>> fd_1.ndim_image
1
```

Iterate through the image

```
>>> for fd_i in fd.image():
...     do_stuff(fd_i)
```

Reviewers

 manso92

 vnmabus

Assignees

No one—assign yourself

Labels

None yet

Projects

None yet

Milestone

No milestone

Desarrollo

- Código público en *Github*
- Estándares de codificación
- Trabajo en equipo
- Documentación
- Pruebas
- Integración continua

skfda.preprocessing.registration.shift_registration

```
skfda.preprocessing.registration.shift_registration(fd, *, maxiter=5, tol=0.01,  
restrict_domain=False, extrapolation=None, step_size=1, initial=None, eval_points=None, **kwargs) \[source\]
```

Perform shift registration of the curves.

Realizes a registration of the curves, using shift alignment, as is defined in [\[RS05-7-2\]](#).

Calculates δ_i for each sample such that $x_i(t + \delta_i)$ minimizes the least squares criterion:

$$\text{REGSSE} = \sum_{i=1}^N \int_{\mathcal{T}} [x_i(t + \delta_i) - \hat{\mu}(t)]^2 ds$$

Estimates the shift parameter δ_i iteratively by using a modified Newton-Raphson algorithm, updating the mean in each iteration, as is described in detail in [\[RS05-7-9-1\]](#).

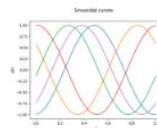
- Parameters:
- **fd** (`FData`) – Functional data object to be registered.
 - **maxiter** (`int`, *optional*) – Maximum number of iterations. Defaults to 5.
 - **tol** (`float`, *optional*) – Tolerance allowable. The process will stop if $\max_i |\delta_i^{(v)} - \delta_i^{(v-1)}| < \text{tol}$. Default sets to $1e-2$.
 - **restrict_domain** (`bool`, *optional*) – If True restricts the domain to avoid evaluate points outside the domain using extrapolation. Defaults uses extrapolation.

Desarrollo

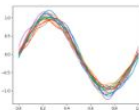
- Código público en *Github*
- Estándares de codificación
- Trabajo en equipo
- Documentación
- Pruebas
- Integración continua

Examples

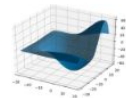
Examples of several functionalities of the package.



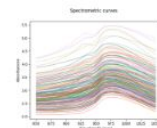
Discretized function
representation



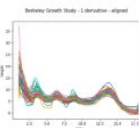
Shift Registration of
basis



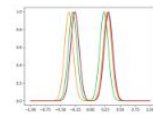
Function composition



Exploring data




Elastic registration



Landmark registration

Desarrollo

- Código público en *Github*
- Estándares de codificación
- Trabajo en equipo
- Documentación
- Pruebas
- Integración continua

GAA-UAM / scikit-fda  build passing

[Current](#) [Branches](#) [Build History](#) [Pull Requests](#)

✓ **Pull Request #112** Feature/neighbors 🔗 #1117 passed

Co-Authored-By: Carlos Ramos Carreño <vnmabus@gmail.com> ⌚ Ran for 6 min 36 sec

🔗 Commit ce77743 🕒 Total time 9 min 4 sec

🔗 #112: Feature/neighbors 📅 about 16 hours ago

🔗 Branch develop

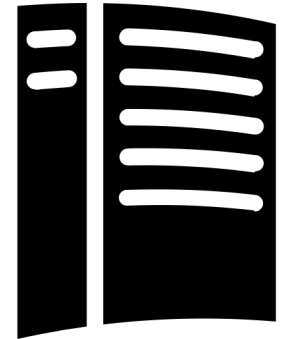
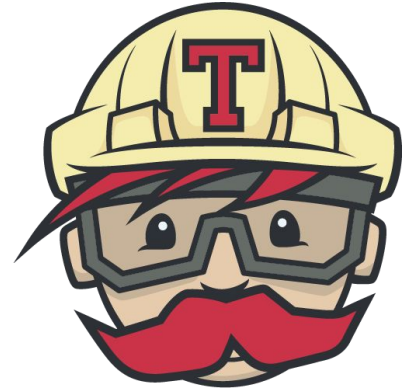
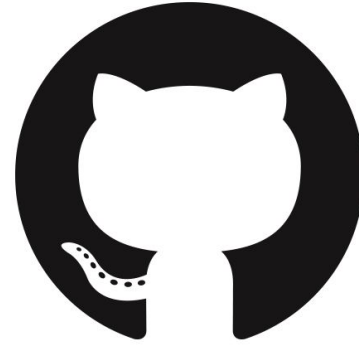
👤 Pablo Marcos

[Build jobs](#) [View config](#)

✓ # 1117.1	🔗 Python 3.6 on Linux
✓ # 1117.2	🔗 Python 3.7.1 on Xenial Linux
✓ # 1117.3	🔗 Python 3.7.2 on macOS
✓ # 1117.4	🔗 Python 3.7.3 on Windows
✓ # 1117.5	🔗 Coverage and pep 8 tests on Python 3.7.1 on Xenial Linux

Desarrollo

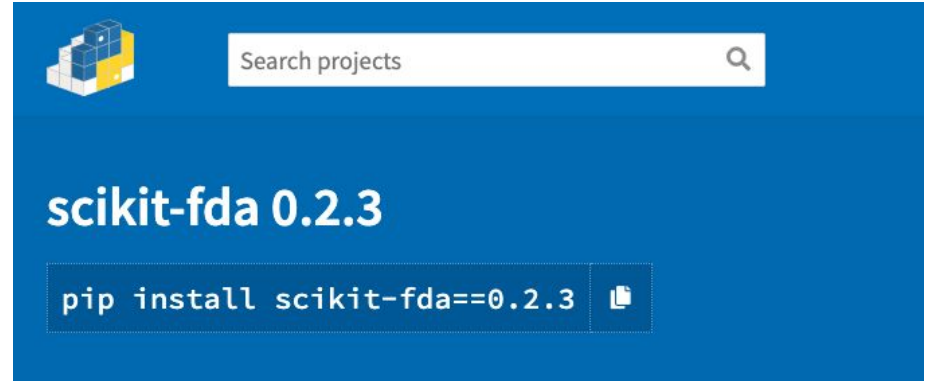
- Código público en *Github*
- Estándares de codificación
- Trabajo en equipo
- Documentación
- Pruebas
- Integración continua





Desarrollo

- Código público en *Github*
- Estándares de codificación
- Trabajo en equipo
- Documentación
- Pruebas
- Integración continua



Representación	Exploración	Preprocesamiento	Inferencia	Machine learning
Paramétrica <ul style="list-style-type: none"> Bases No paramétrica <ul style="list-style-type: none"> Densa Dispersa 	Visualización Estadísticos Profundidad Datos atípicos Reducción de dimensionalidad Distancias	Suavizado Registro Derivación Transformaciones	Tests Intervalos de confianza	Clasificación Regresión Clustering

Representación	<ul style="list-style-type: none">• Diseño unificado API• Ampliación métodos de las representaciones• Optimización representación en bases• API para la evaluación, interpolación y extrapolación
Exploración	<ul style="list-style-type: none">• Creación de módulo con métricas• Métricas Lp multivariantes y elásticas
Preprocesamiento	<ul style="list-style-type: none">• API para el registro de datos• Transformaciones para análisis elástico
Aprendizaje automático	<ul style="list-style-type: none">• Estimadores de vecinos próximos, regresión y clasificación
General	<ul style="list-style-type: none">• Mejoras documentación paquete• Ampliación bancos de pruebas existentes• Configuración de herramientas CI

Análisis	<ul style="list-style-type: none">• Análisis funcional• Variable real• Análisis matemático
Geometría	<ul style="list-style-type: none">• Geometría de curvas y superficies• Geometría diferencial
Estadística	<ul style="list-style-type: none">• Estadística• Aprendizaje automático
Algoritmos	<ul style="list-style-type: none">• Cálculo numérico• Programación• Análisis de algoritmos
Software	<ul style="list-style-type: none">• Ingeniería del software• Análisis y diseño de software

iMuchas gracias!