Bagplots, boxplots and outlier detection for functional data

Han Lin Shang & Rob J Hyndman

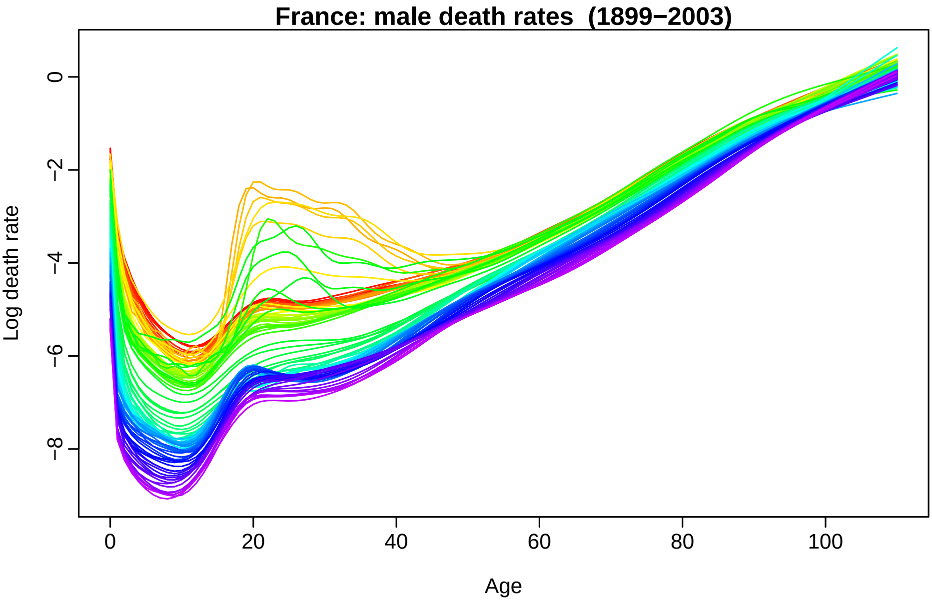
Business & Economic Forecasting Unit

MONASH University

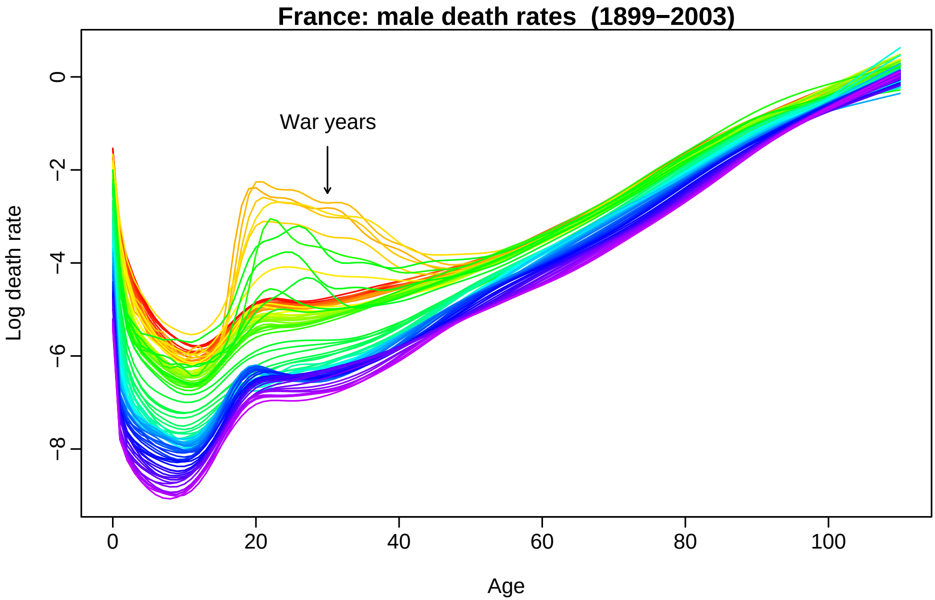
Outline

Outline

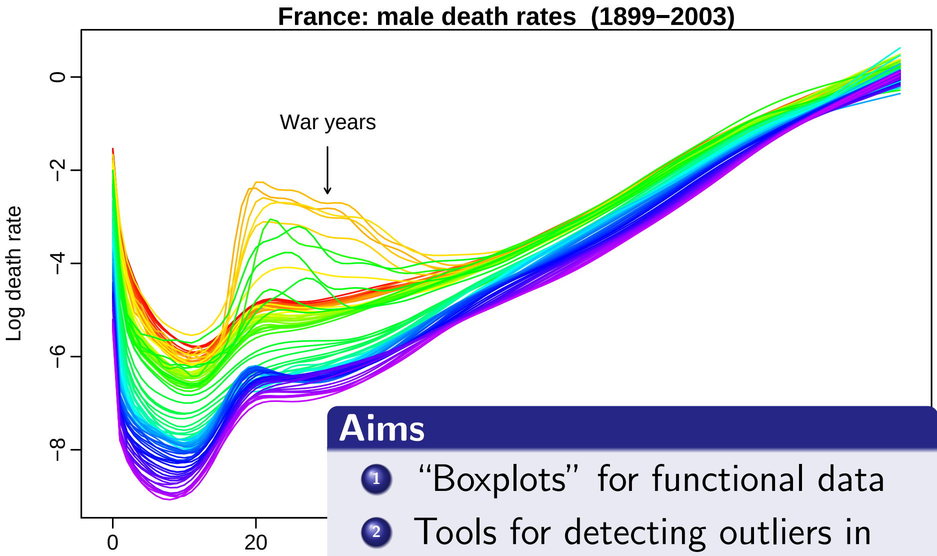
French male mortality rates



French male mortality rates



French male mortality rates



Aims

1. "Boxplots" for functional data
2. Tools for detecting outliers in functional data

Robust principal components

Let $\{y_i(x)\}$, $i = 1, \ldots, n$, be a set of curves.

Robust principal components

Let $\{y_i(x)\}$, $i = 1, \dots, n$, be a set of curves.

1. **Apply a robust principal component algorithm**

$$y_i(x_i) = \mu(x) + \sum_{k=1}^{n-1} z_{i,k}\phi_k(x)$$

Robust principal components

Let $\{y_i(x)\}$, $i = 1, \dots, n$, be a set of curves.

1. **Apply a robust principal component algorithm**

$$y_i(x_i) = \mu(x) + \sum_{k=1}^{n-1} z_{i,k}\phi_k(x)$$

- $\mu(x)$ is mean curve

Robust principal components

Let $\{y_i(x)\}$, $i = 1, \dots, n$, be a set of curves.

1. **Apply a robust principal component algorithm**

$$y_i(x_i) = \mu(x) + \sum_{k=1}^{n-1} z_{i,k}\phi_k(x)$$

- $\mu(x)$ is mean curve
- $\{\phi_k(x)\}$ are principal components

Robust principal components

Let $\{y_i(x)\}$, $i = 1, \dots, n$, be a set of curves.

1. **Apply a robust principal component algorithm**

$$y_i(x_i) = \mu(x) + \sum_{k=1}^{n-1} z_{i,k}\phi_k(x)$$

- $\mu(x)$ is mean curve
- $\{\phi_k(x)\}$ are principal components
- $\{z_{i,k}\}$ are PC scores

Robust principal components

Let $\{y_i(x)\}$, $i = 1, \dots, n$, be a set of curves.

1. **Apply a robust principal component algorithm**

$$y_i(x_i) = \mu(x) + \sum_{k=1}^{n-1} z_{i,k}\phi_k(x)$$

   - $\mu(x)$ is mean curve
   - $\{\phi_k(x)\}$ are principal components
   - $\{z_{i,k}\}$ are PC scores

2. **Plot $z_{i,2}$ vs $z_{i,1}$**

Robust principal components

Let $\{y_i(x)\}$, $i = 1, \dots, n$, be a set of curves.

1. **Apply a robust principal component algorithm**

$$y_i(x_i) = \mu(x) + \sum_{k=1}^{n-1} z_{i,k}\phi_k(x)$$

   - $\mu(x)$ is mean curve
   - $\{\phi_k(x)\}$ are principal components
   - $\{z_{i,k}\}$ are PC scores

2. **Plot $z_{i,2}$ vs $z_{i,1}$**

Robust principal components

Let $\{y_i(x)\}$, $i = 1, \dots, n$, be a set of curves.

1. **Apply a robust principal component algorithm**

$$y_i(x_i) = \mu(x) + \sum_{k=1}^{n-1} z_{i,k}\phi_k(x)$$

   - $\mu(x)$ is mean curve
   - $\{\phi_k(x)\}$ are principal components
   - $\{z_{i,k}\}$ are PC scores

2. **Plot $z_{i,2}$ vs $z_{i,1}$**

➥ Each point in scatterplot represents one curve.

Robust principal components

Let $\{y_i(x)\}$, $i = 1, \ldots, n$, be a set of curves.

1. **Apply a robust principal component algorithm**

$$y_i(x_i) = \mu(x) + \sum_{k=1}^{n-1} z_{i,k}\phi_k(x)$$

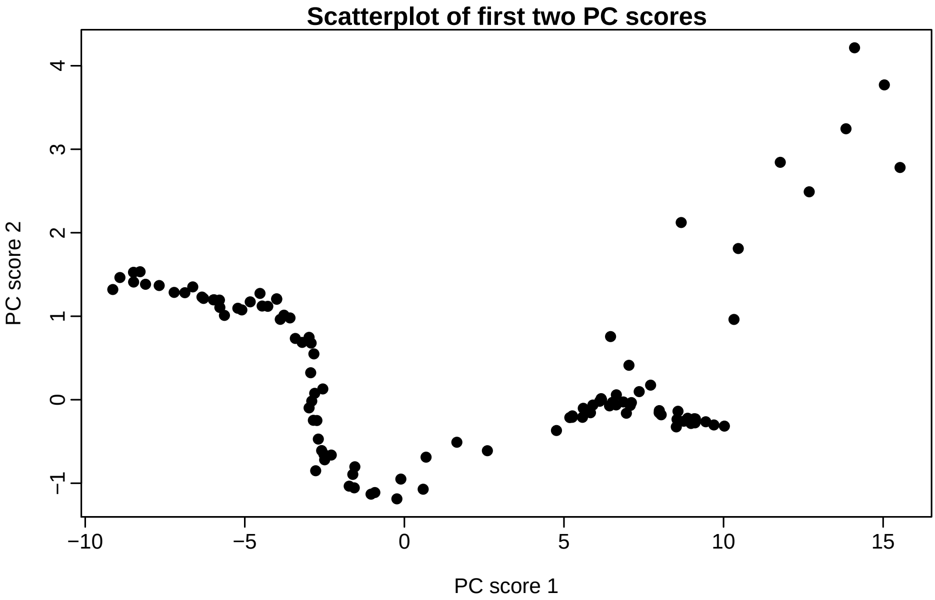
- $\mu(x)$ is mean curve
- $\{\phi_k(x)\}$ are principal components
- $\{z_{i,k}\}$ are PC scores

2. **Plot $z_{i,2}$ vs $z_{i,1}$**

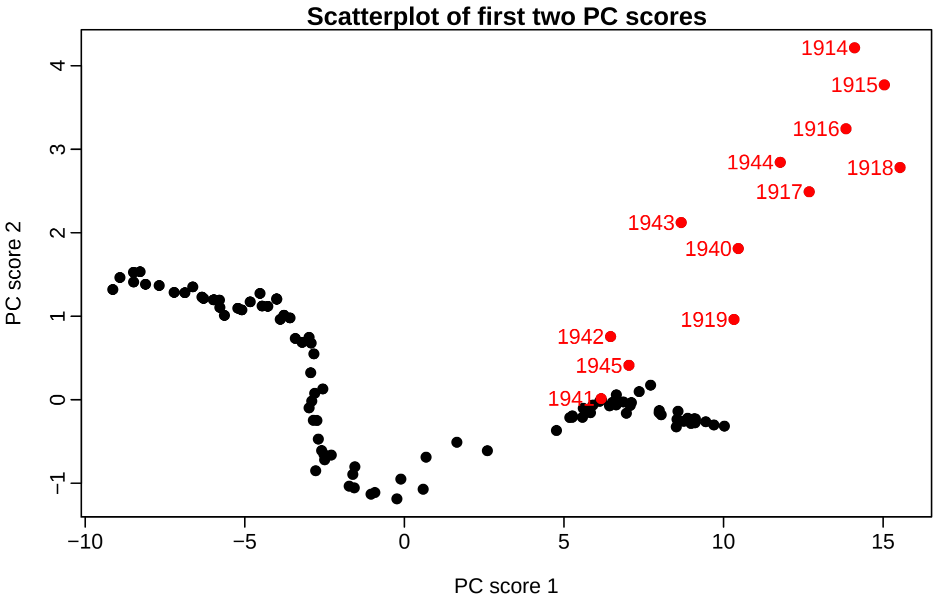
➥ Each point in scatterplot represents one curve.

➥ Outliers show up in bivariate score space.

Robust principal components



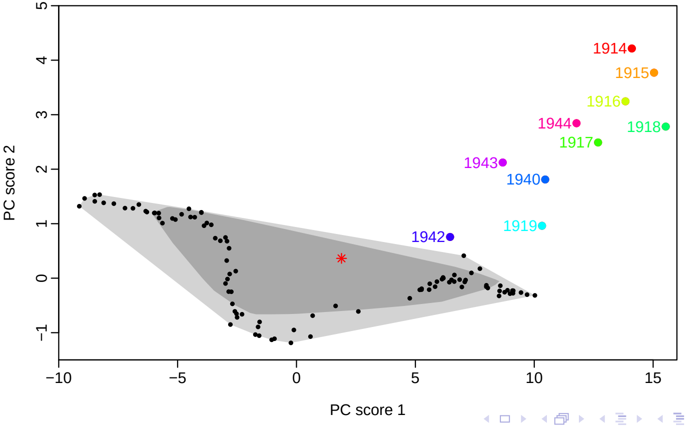
Robust principal components



Outline

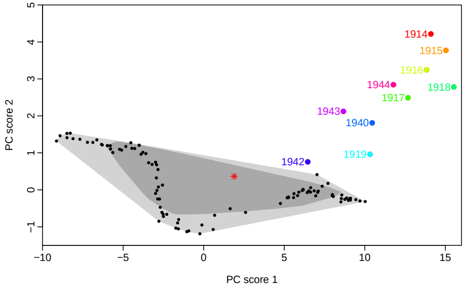
Functional bagplot

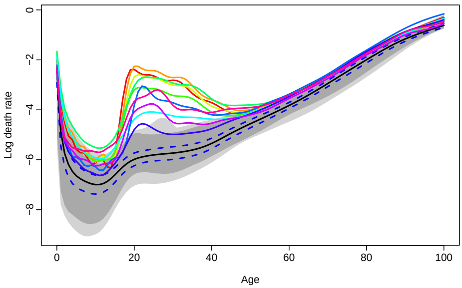
- Bivariate bagplot due to Rousseeuw et al. (1999).
- Rank points by halfspace location depth.
- Display median, 50% convex hull and outer convex hull (with 99% coverage if bivariate normal).



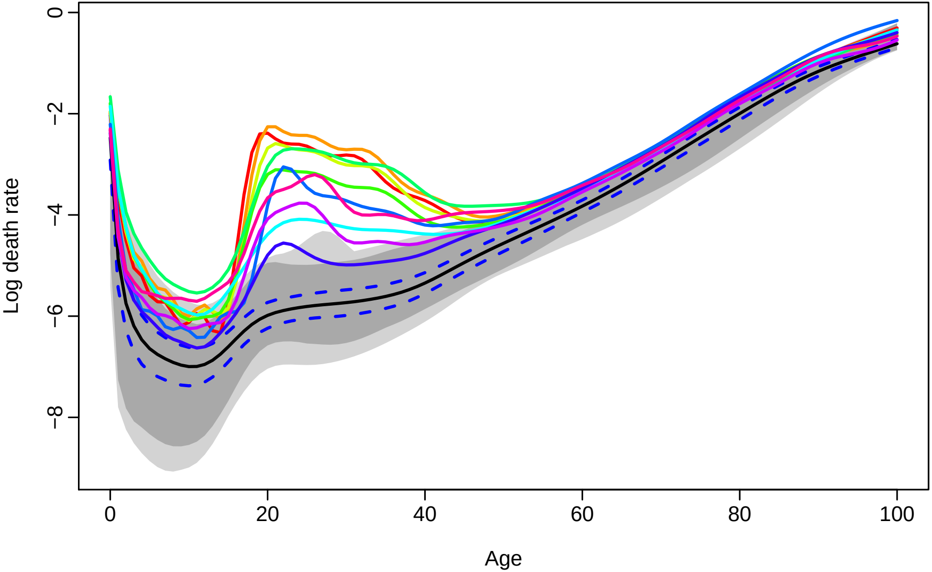
Functional bagplot

- Bivariate bagplot due to Rousseeuw et al. (1999).
- Rank points by halfspace location depth.
- Display median, 50% convex hull and outer convex hull (with 99% coverage if bivariate normal).
- Boundaries contain all curves inside bags.
- 95% CI for median curve also shown.



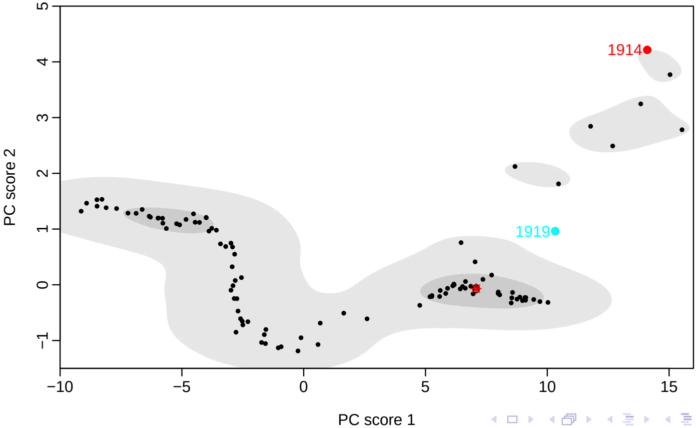


Functional bagplot



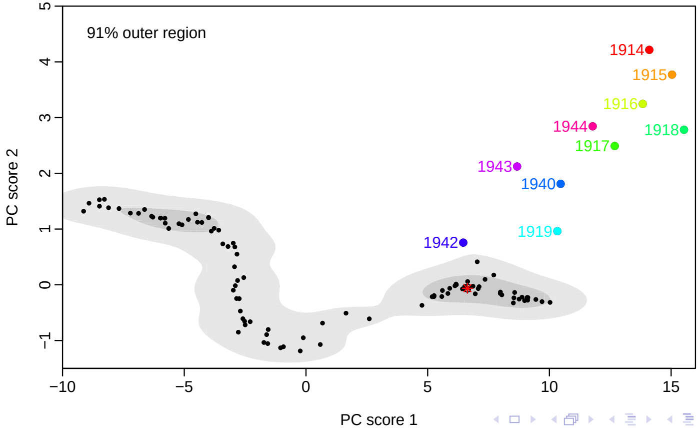
Functional HDR boxplot

- Bivariate HDR boxplot due to Hyndman (1996).
- Rank points by value of kernel density estimate.
- Display mode, 50% and (usually) 99% highest density regions (HDRs) and mode.



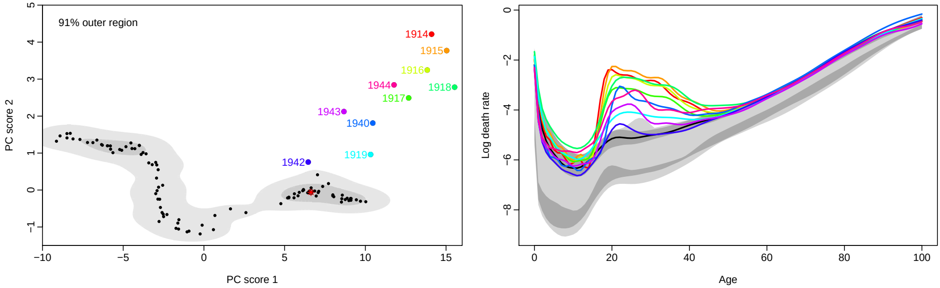
Functional HDR boxplot

- Bivariate HDR boxplot due to Hyndman (1996).
- Rank points by value of kernel density estimate.
- Display mode, 50% and (usually) 99% highest density regions (HDRs) and mode.

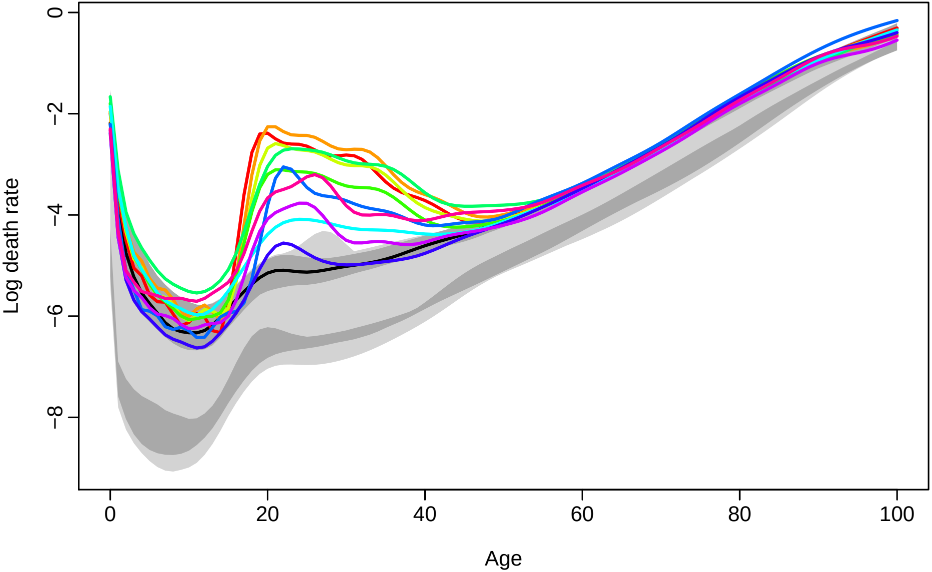


Functional HDR boxplot

- Bivariate HDR boxplot due to Hyndman (1996).
- Rank points by value of kernel density estimate.
- Display mode, 50% and (usually) 99% highest density regions (HDRs) and mode.
- Boundaries contain all curves inside HDRs.



Functional HDR boxplot



Outline

Outlier detection: existing methods

Likelihood ratio method

- Febrero et al. (2007) find curve that maximizes LRT statistic.
- If LRT $> C$, then curve is considered outlier.
- C is computed via smoothed bootstrap.
- Process continues until no more outliers.

Outlier detection: existing methods

Likelihood ratio method

- Febrero et al. (2007) find curve that maximizes LRT statistic.
- If LRT $> C$, then curve is considered outlier.
- C is computed via smoothed bootstrap.
- Process continues until no more outliers.

Disadvantages

- Computationally intensive.
- Ignores shape outliers.
- If trimmed mean is used and there is no outlier, C will be downward biased.

Outlier detection: existing methods

Integrated squared error method

- Hyndman & Ullah (2007) proposed the use of

$$v_i = \int_x \left[\hat{y}_i(x) - \mu(x) - \sum_{k=1}^{K} z_{i,k}\phi_k(x)\right]^2 dx$$

where $z_{i,k}$ and (robust) PC scores and $\phi_k(x)$ are PCs.

Outlier detection: existing methods

Integrated squared error method

- Hyndman & Ullah (2007) proposed the use of

$$v_i = \int_x \left[\hat{y}_i(x) - \mu(x) - \sum_{k=1}^{K} z_{i,k}\phi_k(x)\right]^2 dx$$

  where $z_{i,k}$ and (robust) PC scores and $\phi_k(x)$ are PCs.
- Curve is outlier if $v_i > s + \lambda\sqrt{s}$, where $s = \text{median}(v_1, \cdots, v_t)$ and λ is tuning parameter.

Outlier detection: existing methods

Integrated squared error method

- Hyndman & Ullah (2007) proposed the use of

$$v_i = \int_x \left[\hat{y}_i(x) - \mu(x) - \sum_{k=1}^{K} z_{i,k}\phi_k(x)\right]^2 dx$$

  where $z_{i,k}$ and (robust) PC scores and $\phi_k(x)$ are PCs.
- Curve is outlier if $v_i > s + \lambda\sqrt{s}$, where $s = \text{median}(v_1, \cdots, v_t)$ and λ is tuning parameter.

Outlier detection: existing methods

Integrated squared error method

- Hyndman & Ullah (2007) proposed the use of

$$v_i = \int_x \left[\hat{y}_i(x) - \mu(x) - \sum_{k=1}^{K} z_{i,k}\phi_k(x) \right]^2 dx$$

  where $z_{i,k}$ and (robust) PC scores and $\phi_k(x)$ are PCs.
- Curve is outlier if $v_i > s + \lambda\sqrt{s}$, where $s = \text{median}(v_1, \cdots, v_t)$ and λ is tuning parameter.

Disadvantages

- Depends on K and λ.
- If K large, outliers modelled by higher components.

Outlier detection: comparison

French male mortality data set

Based on historical information, the outliers are expected to be 1914–1919 & 1940–1945.

Method	Outliers detected
Likelihood ratio	—
Integrated squared error	1914–1918, 1940, 1943–1944
Bagplot	1914–1919, 1940, 1942–1944
91% HDR boxplot	1914–1919, 1940, 1942–1944

Outlier detection: comparison

French male mortality data set

Based on historical information, the outliers are expected to be 1914–1919 & 1940–1945.

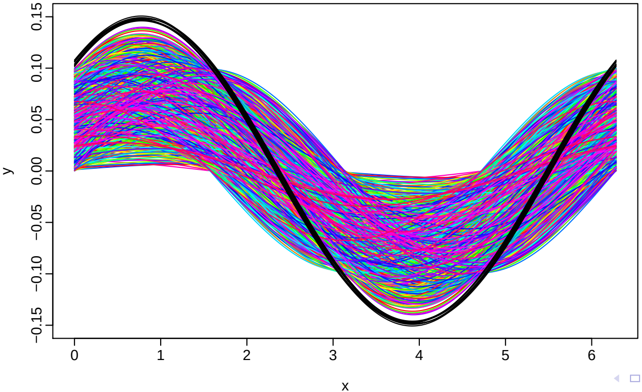
Method	Outliers detected
Likelihood ratio	—
Integrated squared error	1914–1918, 1940, 1943–1944
Bagplot	1914–1919, 1940, 1942–1944
91% HDR boxplot	1914–1919, 1940, 1942–1944

Method	Sensitivity	Specificity	Time (s)
Likelihood ratio	0%	100%	18.8
Integrated squared error	50%	94%	3.4
Bagplot	83%	98%	0.6
91% HDR boxplot	83%	98%	0.3

Outlier detection: comparison

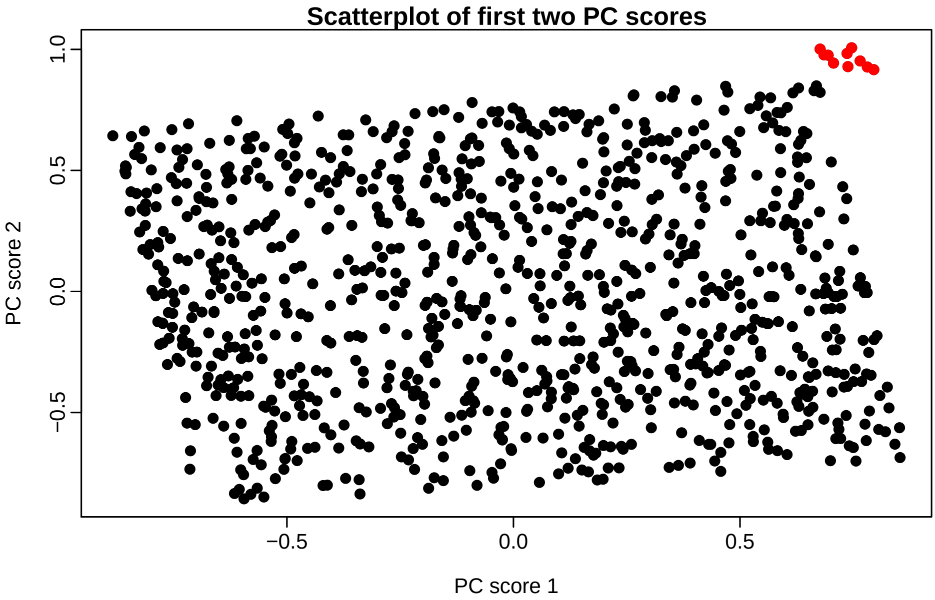
Simulation

- $y_i(x) = a_i \sin(x) + b_i \cos(x), \qquad 0 < x < 2\pi$
- $a_i, b_i \sim \mathrm{Unif}(0, 0.1)$ with probability 99%
- $a_i, b_i \sim \mathrm{Unif}(0.1, 0.108)$ with probability 1%



Outliers shown in black

Outlier detection: comparison



Outlier detection: comparison

Simulation

Method	Outliers detected
Likelihood ratio	—
Integrated squared error	—
Bagplot	—
99% HDR boxplot	All

Outlier detection: comparison

Simulation

Method	Outliers detected
Likelihood ratio	—
Integrated squared error	—
Bagplot	—
99% HDR boxplot	All

Method	Sensitivity	Specificity	Time (s)
Likelihood ratio	0%	100%	28.5
Integrated squared error	0%	100%	18.8
Bagplot	0%	100%	7.3
99% HDR boxplot	100%	100%	6.9

Outline

Conclusions

- Functional bagplot highly robust but sometimes misses outliers.

Conclusions

- Functional bagplot highly robust but sometimes misses outliers.
- Functional HDR boxplot more flexible but coverage probability needs tuning.

Conclusions

- Functional bagplot highly robust but sometimes misses outliers.
- Functional HDR boxplot more flexible but coverage probability needs tuning.
- Functional HDR boxplot can detect bimodality and inliers.

Conclusions

- Functional bagplot highly robust but sometimes misses outliers.
- Functional HDR boxplot more flexible but coverage probability needs tuning.
- Functional HDR boxplot can detect bimodality and inliers.
- Existing depth method performs poorly and ignores shape outliers.

Conclusions

- Functional bagplot highly robust but sometimes misses outliers.
- Functional HDR boxplot more flexible but coverage probability needs tuning.
- Functional HDR boxplot can detect bimodality and inliers.
- Existing depth method performs poorly and ignores shape outliers.
- Existing ISE method often misses outliers.

Conclusions

- Functional bagplot highly robust but sometimes misses outliers.
- Functional HDR boxplot more flexible but coverage probability needs tuning.
- Functional HDR boxplot can detect bimodality and inliers.
- Existing depth method performs poorly and ignores shape outliers.
- Existing ISE method often misses outliers.

Conclusions

- Functional bagplot highly robust but sometimes misses outliers.
- Functional HDR boxplot more flexible but coverage probability needs tuning.
- Functional HDR boxplot can detect bimodality and inliers.
- Existing depth method performs poorly and ignores shape outliers.
- Existing ISE method often misses outliers.

➥ Paper and R code: **www.robhyndman.info**

➥ Comments to: **Han.Shang@buseco.monash.edu**