Sparse Functional Boxplots for Multivariate Curves

Zhuo Qu, Marc G. Genton

Statistics Program, King Abdullah University of Science and Technology, Saudi Arabia

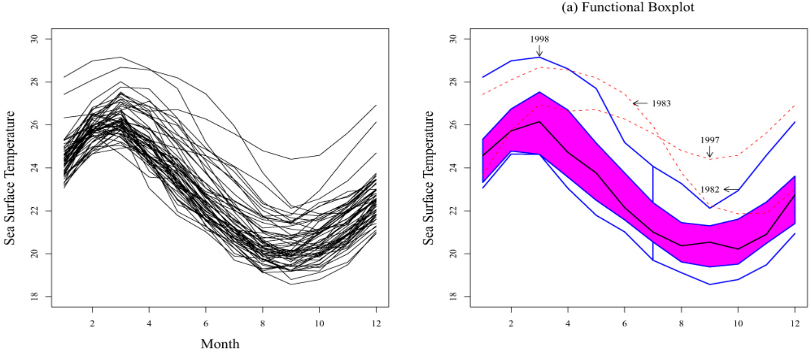
October 26, 2022

Overview: Sparse Functional Boxplots for Multivariate Curves

1. Motivation
   - Restrictions of current visualization tools: measured on common grids
   - Limited discussion of depths for sparse multivariate functional data
2. Preparatory Work
   - Fitting sparse multivariate functional data
   - Depths for sparse multivariate functional data
3. Construction of Sparse Visualization Tools
   - Sparse functional boxplot
   - Intensity sparse functional boxplot
4. Application: Malnutrition Data

1. Motivation I: Restriction of the current visualization tool

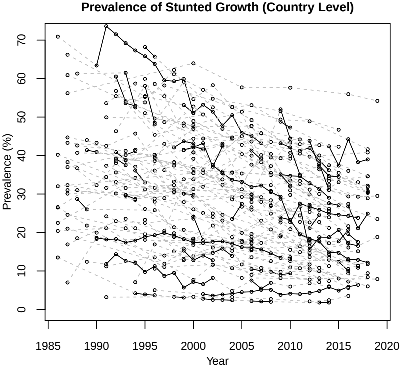
1. Functional Boxplot (Sun & Genton 2011) is an exploratory visualization tool for functional data



2. Enhanced or adjusted functional boxplots (Sun & Genton 2012), surface boxplot (Genton et al. 2014), two-stage functional boxplot (Dai & Genton 2018), trajectory functional boxplot (Yao et al. 2020)

Motivation: Malnutrition Example

How about data with missng values? (Apply them to our tool: Figure 5)



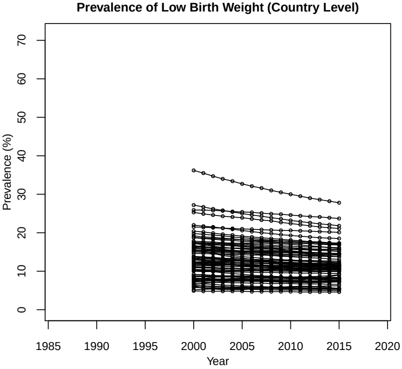


Figure: The observed prevalence of stunted growth and prevalence of low birth weight for 77 countries from 1985 to 2019 (source). Observations are joined with solid black lines if observed continuously; otherwise, joined with gray dashed lines.

Motivation II: Limited discussion of ordering sparse multivariate functional data

- Visualization tool for sparse multivariate functional data

$$\Downarrow$$

  Ranking sparse multivariate functional data

- Current methods to order multivariate functional data:

  1. Depth-based: weighted average of marginal functional depth (WMFD, Ieva & Paganoni 2013), modified simplicial band depth (MSBD, López-Pintado et al. 2014), multivariate functional halfspace depth (MFHD, Claeskens et al. 2014)
  2. Outlying-based: directional outlyingness (DO, Dai & Genton 2019)

Motivations

Restrictions:

1. Visualization tools for univariate (multivariate) functional data $\searrow$
2. Methods for ordering multivariate functional data $\swarrow$

assume that data are observed on common grids

Thread:

- Depths for sparse multivariate functional data ($\times$, not considered)
- Depths for univarariate sparse functional data ($\checkmark$, López-Pintado & Wei (2011) and Sguera & López-Pintado (2020))

Contributions

Logic:

1. Impute missing values
2. Propose multivariate functional depth
3. Add features of sparseness in functional boxplot

Contributions:

1. improve sparse multivariate functional data fitting
2. consider possible depths for sparse multivariate functional data
3. propose exploratory visualization tools for both univariate and multivariate functional data with missing values

2. Preparatory Work

To propose the depth for sparse multivariate functional data, recall:

I. Happ & Greven (2018) proposed the principal component analysis for multivariate functional data (MFPCA) that are defined on different time domains

$\Downarrow$

I. explore the improvement of sparse multivariate data fitting

II. López-Pintado & Wei (2011) and Sguera & López-Pintado (2020) proposed possible depths for univariate sparse functional data

$\Downarrow$

II. consider several depths to order sparse multivariate functional data

Preparatory Work I: Data Fitting

Data Notation:

- $\mathcal{T} = \mathcal{T}_1 \times \cdots \times \mathcal{T}_p$, where $(\mathcal{T}_j)_{j=1}^p$ is a compact set in $\mathbb{R}^{d_j}$
- $\boldsymbol{Y}(\boldsymbol{t}) = (Y^{(1)}(t^{(1)}), \ldots, Y^{(p)}(t^{(p)}))^\top$, $\boldsymbol{t} := (t^{(1)}, \ldots, t^{(p)}) \in \mathcal{T}$
- $Y^{(j)}(t^{(j)}) : \mathcal{T}_j \to \mathbb{R}$ is assumed to be square-integrable in $\mathcal{T}_j$
- $\widetilde{\boldsymbol{Y}}(\boldsymbol{t}) = (\widetilde{Y}^{(1)}(t^{(1)}), \ldots, \widetilde{Y}^{(p)}(t^{(p)}))^\top \in \mathbb{R}^p$, due to measurement errors $\boldsymbol{\epsilon}_i = (\epsilon_i^{(1)}, \ldots, \epsilon_i^{(p)})^\top$ with $\epsilon_i^{(j)} \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_j^2)$

The ith $(i = 1, \ldots, N)$ observation $\widetilde{\boldsymbol{Y}}_i(\boldsymbol{t}_i)$ is denoted as

$$\widetilde{\boldsymbol{Y}}_i(\boldsymbol{t}_i) = \boldsymbol{Y}_i(\boldsymbol{t}_i) + \boldsymbol{\epsilon}_i = \boldsymbol{\mu}(\boldsymbol{t}_i) + \sum_{m=1}^{\infty} \rho_{i,m} \boldsymbol{\psi}_{i,m}(\boldsymbol{t}_i) + \boldsymbol{\epsilon}_i,$$

where $\boldsymbol{\mu}(\boldsymbol{t}_i)$ is the mean function, $\boldsymbol{\psi}_{i,m}(\boldsymbol{t}_i)$ and $\rho_{i,m}$ are the multivariate eigenfunctions and eigenscores respectively

Preparatory Work I: Data Fitting

Happ & Greven (2018) proposed the MFPCA fit based on unobserved MFPC decomposition objects $\boldsymbol{\theta} = \{\boldsymbol{\mu}, \rho, \boldsymbol{\psi}\}$:

$$\widehat{\boldsymbol{Y}}_{\widehat{\boldsymbol{\theta}},i} := \mathrm{E}\left[\widetilde{\boldsymbol{Y}}_i | \widehat{\boldsymbol{\theta}}\right] = \widehat{\boldsymbol{\mu}}_i + \sum_{m=1}^{M} \widehat{\rho}_{i,m} \widehat{\boldsymbol{\psi}}_m, \quad i = 1, \dots, N$$

1. Due to the uncertainty of the obtained eigenvalues and eigenfunctions (Goldsmith et al, 2013), we propose the bootstrap improved MFPCA fit (BMFPCA fit):

$$\widehat{\boldsymbol{Y}}_i = \mathrm{E}_{\widehat{\boldsymbol{\theta}}}\left\{\mathrm{E}_{\widetilde{\boldsymbol{Y}}_i|\widehat{\boldsymbol{\theta}}}\left[\widetilde{\boldsymbol{Y}}_i | \widehat{\boldsymbol{\theta}}\right]\right\}$$

2. $(1-\alpha)$ bootstrap confidence band: we take the $(1-\frac{\alpha}{2})$th and the $\frac{\alpha}{2}$th percentiles across the bootstrap as the confidence upper bound $\widehat{\boldsymbol{Y}}_{ub,i}$ and lower bound $\widehat{\boldsymbol{Y}}_{lb,i}$, respectively

Preparatory Work II: Possible Depth Notions

Take multivariate functional halfspace depth (MFHD) as an example

$$MFHD(\boldsymbol{X}; F_{\boldsymbol{\mathcal{Y}}}, \beta) = \int_{\mathcal{T}} HD(\boldsymbol{X}(t); F_{\boldsymbol{\mathcal{Y}}(t)}) \cdot w_{\beta}(t; F_{\boldsymbol{\mathcal{Y}}(\boldsymbol{t})})\mathrm{d}t$$

Note:

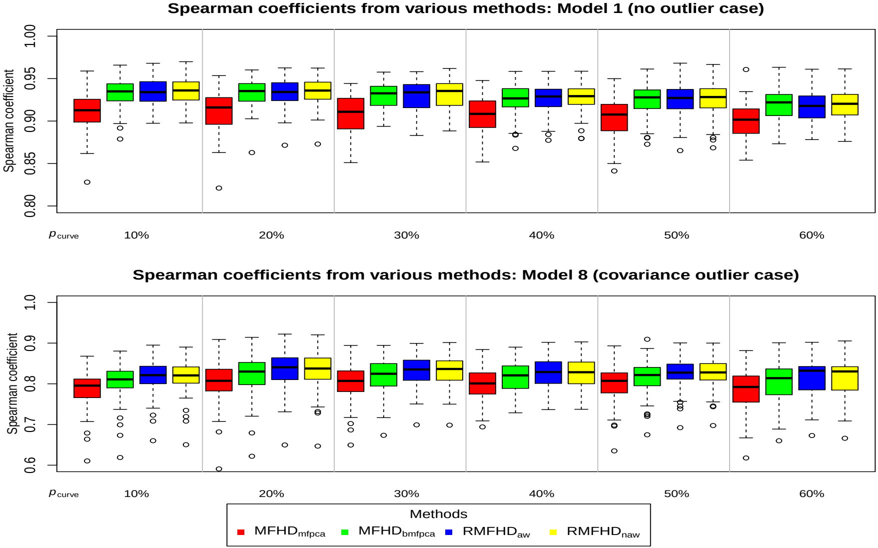
1. $HD(\boldsymbol{X}(t); F_{\boldsymbol{\mathcal{Y}}(t)})$: $\mathbb{R}^p \mapsto \mathbb{R}$ is the halfspace depth of $\boldsymbol{X}(t)$ with respect to the random variable with cumulative distribution function (cdf) $F_{\boldsymbol{\mathcal{Y}}(t)}$
2. $w_{\beta}(t; F_{\boldsymbol{\mathcal{Y}}(t)})$ is a weight function satisfying $\int_{\mathcal{T}} w_{\beta}(t; F_{\boldsymbol{\mathcal{Y}}(t)})\mathrm{d}t = 1$ for a fixed $\beta \in (0, 1]$

Preparatory Work II: Possible Depth Notions

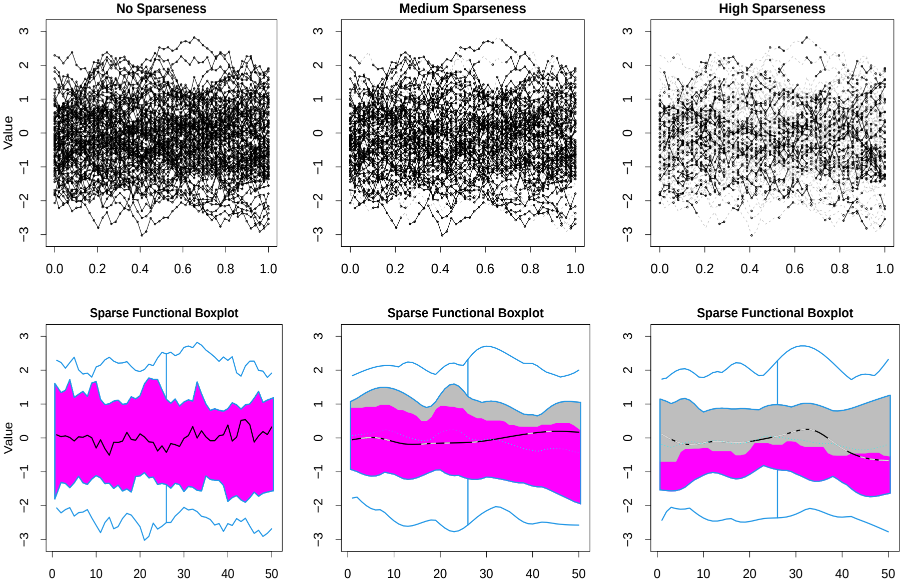
Let $\widehat{\boldsymbol{Y}}^f = \{\widehat{\boldsymbol{Y}}_1, \ldots, \widehat{\boldsymbol{Y}}_N\}$ be a set of fitted data, $\widehat{\boldsymbol{Y}}^f_{ub} = \{\widehat{\boldsymbol{Y}}_{ub,1}, \ldots, \widehat{\boldsymbol{Y}}_{ub,N}\}$ be a set of confidence upper bounds, $\widehat{\boldsymbol{Y}}^f_{lb} = \{\widehat{\boldsymbol{Y}}_{lb,1}, \ldots, \widehat{\boldsymbol{Y}}_{lb,N}\}$ be a set of confidence lower bounds, and $\widehat{\boldsymbol{Y}}_{upd} = \{\widehat{\boldsymbol{Y}}^f, \widehat{\boldsymbol{Y}}^f_{ub}, \widehat{\boldsymbol{Y}}^f_{lb}\}$

1. Apply MFHD to each element in $\widehat{\boldsymbol{Y}}^f$: $MFHD(\widehat{\boldsymbol{Y}}; \widehat{\boldsymbol{Y}}^f)$
2. $$\mathrm{RMFHD}_{\mathrm{type}}(\widehat{\mathbf{Y}}; \widehat{\mathbf{Y}}_{upd}) = \begin{cases} \frac{1}{3}\mathrm{MFHD}(\widehat{\mathbf{Y}}; \widehat{\mathbf{Y}}_{upd}) + \frac{1}{3}\mathrm{MFHD}(\widehat{\mathbf{Y}}_{ub}; \widehat{\mathbf{Y}}_{upd}) \\ \quad + \frac{1}{3}\mathrm{MFHD}(\widehat{\mathbf{Y}}_{lb}; \widehat{\mathbf{Y}}_{upd}), & \text{if type} = \text{aw}, \\ \frac{1}{2}\mathrm{MFHD}(\widehat{\mathbf{Y}}; \widehat{\mathbf{Y}}_{upd}) + \frac{1}{4}\mathrm{MFHD}(\widehat{\mathbf{Y}}_{ub}; \widehat{\mathbf{Y}}_{upd}) \\ \quad + \frac{1}{4}\mathrm{MFHD}(\widehat{\mathbf{Y}}_{lb}; \widehat{\mathbf{Y}}_{upd}), & \text{if type} = \text{naw}. \end{cases}$$

Optimal Depth: $\mathrm{MFHD}_{\mathrm{bmfpca}}$



Examples of the Sparse Functional Boxplot



3. Construction of Sparse Visualization Tools

- Let $Z_{[r]}^{(j)}(t)$ be the jth component of the rth deepest curve evaluated at $t \in \mathcal{T}$, and $\lceil n/2 \rceil$ be the smallest integer $\geq n/2$

1. Central region:
$$C_{0.5}^{(j)} := \{(t, Z^{(j)}(t)) : \min_{r=1,\ldots,\lceil n/2 \rceil} Z_{[r]}^{(j)}(t) \leq Z^{(j)}(t) \leq \max_{r=1,\ldots,\lceil n/2 \rceil} Z_{[r]}^{(j)}(t)\}$$
2. The median: $Z_{\lceil 1 \rceil}^{(j)}(t)$
3. The outliers: A curve $\boldsymbol{Z}_o$ is classified in S_o ($\boldsymbol{Z}_o \in S_o$), if
$Z_o^{(j)}(t) > Z_{ub,0.5}^{(j)}(t) + 1.5R_{0.5}^{(j)}(t)$
or
$Z_o^{(j)}(t) < Z_{lb,0.5}^{(j)}(t) - 1.5R_{0.5}^{(j)}(t)$
4. The non-outlying maximal (minimal) bound: the non-outlying maximal bound
$Z_{ub}^{(j)}(t) := \max_{\boldsymbol{Z} \in \boldsymbol{Z}^f \setminus S_o} Z^{(j)}(t)$

Construction of Sparse Functional Boxplot

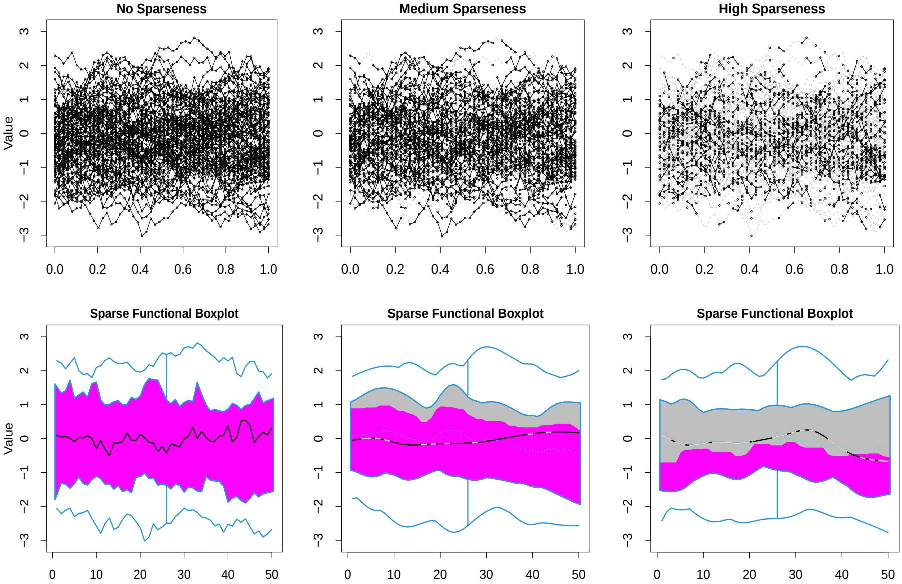
The sparse functional boxplot, apart from displaying the aforementioned features, underlines the sparseness features in the median $Z^{(j)}_{[1]}(t)$, the 50% central region $C^{(j)}_{0.5}$, and the detected outliers S_o

1. The median and outliers: underline the missing values in gray
2. The central region:

- At each time point $t \in \mathcal{T}$, the sparseness proportion $p_s^{(j)}(t) := \frac{n_{ms}^{(j)}(t)}{\{n_{ms}^{(j)}(t)+n_{obs}^{(j)}(t)\}}$ within $C^{(j)}_{0.5}$
- Define the proportion line $l^{(j)}(t, p_s^{(j)}(t)) := Z^{(j)}_{lb,0.5}(t) - p_s^{(j)}(t)R^{(j)}_{0.5}(t)$ for $t \in \mathcal{T}$

Show the observed proportion below the proportion line in magenta and the sparseness proportion above in gray

Examples of the Sparse Functional Boxplot



Construction of Intensity Sparse Functional Boxplot

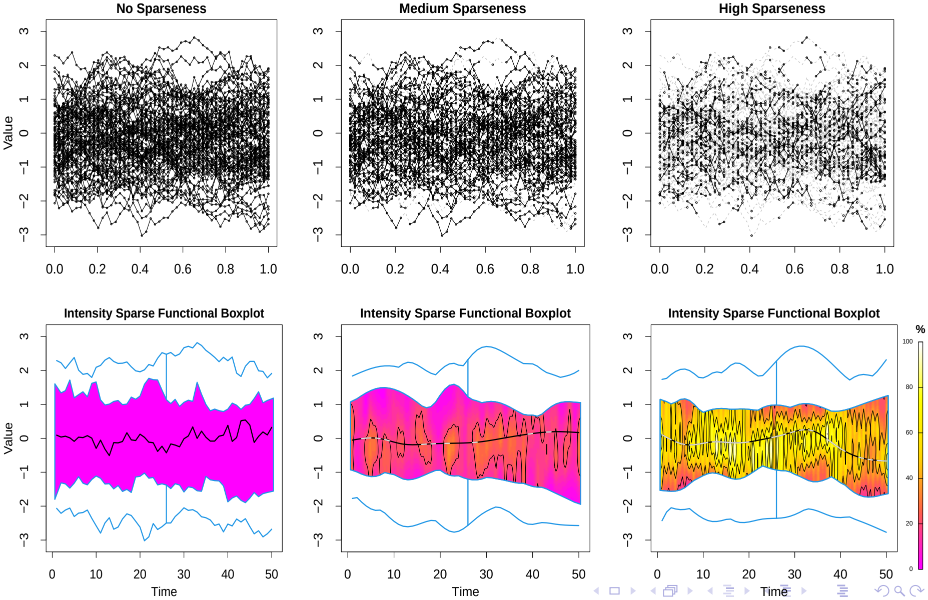
In addition, we display the intensity of fitted missing point patterns within $C_{0.5}^{(j)}$, which is expressed in estimated missing points per unit area

Definition of the Intensity:

- Regard S fitted sparse points within $C_{0.5}^{(j)}$ as a spatial point pattern $\boldsymbol{u}_s^{(j)} := \{(t_s, Z_s^{(j)}) \in C_{0.5}^{(j)}, s = 1, \ldots, S\}$ with t_s the time, and $Z_s^{(j)}$ the fitted value inside the central region
- The sparseness intensity at a new point $\boldsymbol{u}^{(j)} \in C_{0.5}^{(j)}$ is

$$\lambda(\boldsymbol{u}^{(j)}) = e(\boldsymbol{u}^{(j)}) \sum_{s=1}^{S} w_s \mathcal{K}(\boldsymbol{u}_s^{(j)} - \boldsymbol{u}^{(j)}),$$

where $\mathcal{K}$ is the Gaussian smoothing kernel, $e(\boldsymbol{u}^{(j)})$ is an edge correction factor, and w_s is the weight


No Sparseness
Medium Sparseness
High Sparseness
Value
3
2
1
0
−1
−2
−3
0.0
0.2
0.4
0.6
0.8
1.0
Intensity Sparse Functional Boxplot
Intensity Sparse Functional Boxplot
Intensity Sparse Functional Boxplot
0
10
20
30
40
50
Time
%
100
80
60
40
20
0

4. Malnutrition data: Sparse Functional Boxplot

As shown in Figure 1, we have two variables: prevalence of stunted growth, and prevalence of low birth weight from 77 countries from 1985 to 2019, which belong to the point sparseness and partial sparseness, respectively.

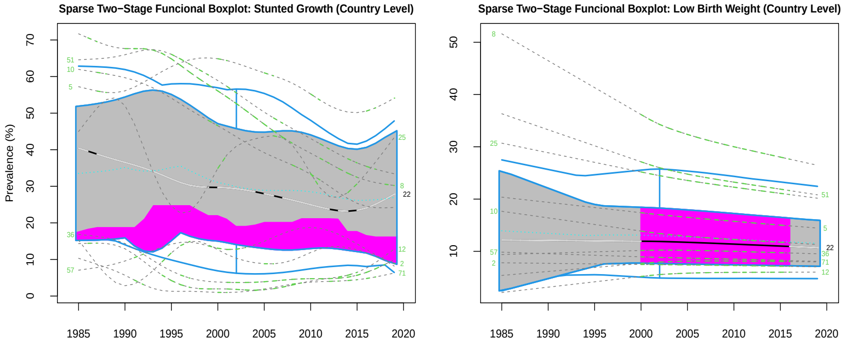


Figure: Visualization of stunted growth and low birth weight data for 77 countries with the sparse two-stage functional boxplot.

Malnutrition data: Intensity Sparse Functional Boxplot

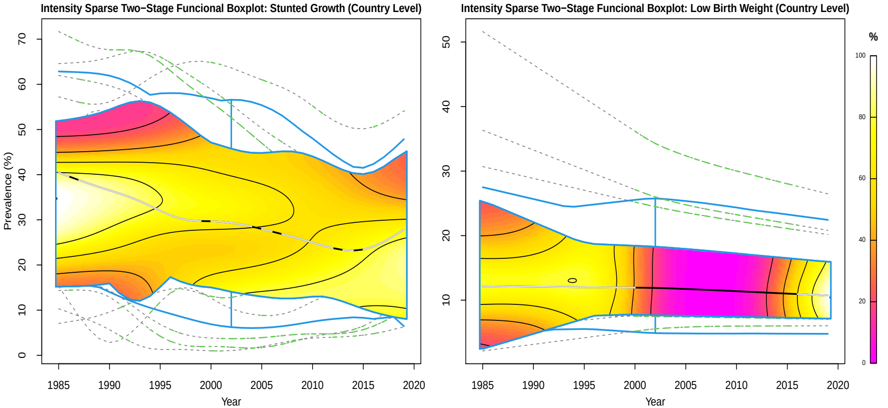


Figure: Visualization of stunted growth and low birth weight data for 77 countries with the intensity sparse two-stage functional boxplot.

Reference and Codes

- Zhuo Qu & Marc G. Genton (2022) Sparse Functional Boxplots for Multivariate Curves, Journal of Computational and Graphical Statistics, DOI: `10.1080/10618600.2022.2066680`
- To replicate the data analysis/apply your data to our visualization tools, please visit: Github

References

Tukey, John W (1975) Mathematics and the picturing of data *Proceedings of the International Congress of Mathematicians, Vancouver* 2, 525 – 531.

López-Pintado, Sara and Romo, Juan (2009) On the concept of depth for functional data *Journal of the American statistical Association* 104 (486), 718 – 734.

Sun, Ying and Genton, Marc G (2011) Functional boxplots *Journal of Computational and Graphical Statistics* 20 (2), 316 – 334.

Sun, Ying and Genton, Marc G (2012) Adjusted functional boxplots for spatio-temporal data visualization and outlier detection *Environmetrics* 23 (1), 54 – 64.

Genton, Marc G and Johnson, Christopher and Potter, Kristin and Stenchikov, Georgiy and Sun, Ying (2014) Surface boxplots *Stat* 3 (1), 1 – 11.

Dai, Wenlin and Genton, Marc G (2018) Functional boxplots for multivariate curves *Stat* 7 (1), e190.

Yao, Zonghui and Dai, Wenlin and G. Genton, Marc (2020) Trajectory functional boxplots *Stat* 9 (1), e289.

Qu, Zhuo and Genton, Marc G (2021) Sparse Functional Boxplots for Multivariate Curves *arXiv* preprint arXiv:2103.07868

Ieva, Francesca and Paganoni, Anna M (2013) Depth measures for multivariate functional data *Communications in Statistics-Theory and Methods* 42 (7), 1265 – 1276.

References

López-Pintado, Sara and Sun, Ying and Lin, Juan K and Genton, Marc G (2014) Simplicial band depth for multivariate functional data *Advances in Data Analysis and Classification* 8 (3), 321 – 338.

Claeskens, Gerda and Hubert, Mia and Slaets, Leen and Vakili, Kaveh (2014) Multivariate functional halfspace depth *Journal of the American Statistical Association* 109 (505), 411 – 423.

Dai, Wenlin and Genton, Marc G (2019) Directional outlyingness for multivariate functional data *Computational Statistics & Data Analysis* 131, 50 – 65.

López-Pintado, Sara and Wei, Ying (2011) Depth for sparse functional data *Recent Advances in Functional Data Analysis and Related Topics* 209 – 212.

Sguera, Carlo and López-Pintado, Sara (2020) A notion of depth for sparse functional data *TEST* 1 – 20.

Happ, Clara and Greven, Sonja (2018) Multivariate functional principal component analysis for data observed on different (dimensional) domains *Journal of the American Statistical Association* 113 (522) 649 – 659.

Goldsmith, Jeff and Greven, Sonja and Crainiceanu, CIPRIAN (2013) Corrected confidence bands for functional data using principal components *Biometrics* 69 (1) 41 – 51.