Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

technische universität dortmund
Mathematical Statistics and Applications in Industry

Two Resampling Approaches for Powerful Inference in Incomplete Paired Functional Data

Marléne Baumeister, Lubna Amro, Łukasz Smaga, Markus Pauly

This work is part of the DFG WAEVE project
Powerful inference for functional data in complex factorial designs

Statistical Question

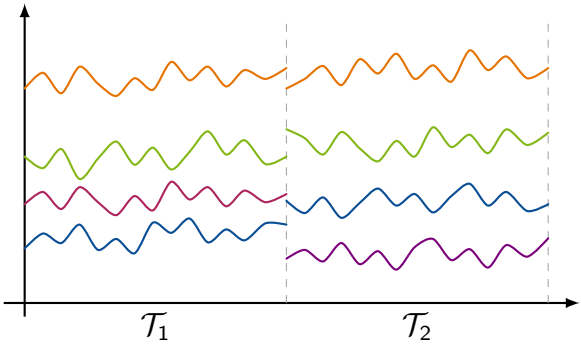
Motivation: Missing values in repeated measurements are unpleasant

- Examples: non-responding in surveys at one time point, lost patients in clinical studies.

$\longrightarrow$ With naive handling, all measured time points must be omitted.

$\longrightarrow$ Many options to handle missing values, mostly based on imputation.

- Handling of missingness in functional data remains challenging.



Missingness in Functional Data

Functional data in a matched pairs design:
Suppose we have a pseudometric compact space $(\mathcal{T}, \rho)$ and $n \in \mathbb{N}$ independent realisations

$$\mathbf{x}_1, \ldots, \mathbf{x}_n : \mathcal{T} \to \mathbb{R}^2$$

of a stochastic process $\mathbf{X}$ and write for every $t \in \mathcal{T}$:

$$\begin{pmatrix} x_{11}(t) \\ x_{21}(t) \end{pmatrix}, \ldots\ldots\ldots, \begin{pmatrix} x_{1n}(t) \\ x_{2n}(t) \end{pmatrix}.$$

Missingness in Functional Data

Functional data in a matched pairs design with missing functions:
Suppose we have a pseudometric compact space $(\mathcal{T}, \rho)$ and $n \in \mathbb{N}$ independent realisations

$$\mathbf{x}_1, \dots, \mathbf{x}_n : \mathcal{T} \to \mathbb{R}^2$$

of a stochastic process $\mathbf{X}$ and write for every $t \in \mathcal{T}$ and $n = n_1 + n_2 + n_3$:

$$\underbrace{\begin{pmatrix} x_{11}^{(c)}(t) \\ x_{21}^{(c)}(t) \end{pmatrix}, \dots, \begin{pmatrix} x_{1n_1}^{(c)}(t) \\ x_{2n_1}^{(c)}(t) \end{pmatrix}}_{\mathbf{x}^{(c)}, \text{ of size } n_1.}, \underbrace{\begin{pmatrix} x_{11}^{(m)}(t) \\ - - - \end{pmatrix}, \dots, \begin{pmatrix} x_{1n_2}^{(m)}(t) \\ - - - \end{pmatrix}, \begin{pmatrix} - - - \\ x_{21}^{(m)}(t) \end{pmatrix}, \dots, \begin{pmatrix} - - - \\ x_{2n_3}^{(m)}(t) \end{pmatrix}}_{\mathbf{x}^{(m)}, \text{ of size } n_2 + n_3.}.$$

Hypothesis

- Functional matched pair hypothesis regarding the pointwise means $\eta_1(t)$ and $\eta_2(t)$, $t \in \mathcal{T}$:

$$\mathcal{H}_0 : \forall\, t \in \mathcal{T} : \eta_1(t) = \eta_2(t), \quad \mathcal{H}_1 : \exists\, t \in \mathcal{T} : \eta_1(t) \neq \eta_2(t). \tag{1}$$

  $\longrightarrow$ Existing methods without missing values: Martínez-Camblor and Corral (2011) and Smaga (2020).

- Non-missing values in $\mathbf{x}^{(m)}$ are i.i.d.,
  $\longrightarrow$ missing observations are missing completely at random (MCAR).
  $\longrightarrow$ Classical imputation methods are not feasible.
- Aim: adapt the approach Amro and Pauly (2017) for the modelling of missing data in matched pairs for functional data.

Statistical Methods

General Set-Up

- Consider stochastic process $\mathbf{X}$ is valued in $C^2(\mathcal{T})$:

$$\{\mathbf{X}(t) = (X_1(t), X_2(t)) : t \in \mathcal{T}\},\ \mathbf{X} : \Omega \to C^2(\mathcal{T}),\ \omega \mapsto \mathbf{X}_t(\omega).$$

  $\longrightarrow$ Similar to the approaches in Dette et al. (2020) and Munko et al. (2025).

- Assume no specific distribution, but mean function $\boldsymbol{\eta}(t) = (\eta_1(t), \eta_2(t))$, $t \in \mathcal{T}$, and covariance matrix

$$\mathbf{\Gamma}(s,t) = \begin{pmatrix} \gamma_1(s,t) & \gamma_{12}(s,t) \\ \gamma_{12}(s,t) & \gamma_2(s,t) \end{pmatrix}, s, t \in \mathcal{T}.$$

Assumptions

1. The covariance matrix $\mathbf{\Gamma}(t,t)$ is positive definite for all $t \in \mathcal{T}$.
2. The complete and incomplete parts do not vanish, i.e. $n_i/(n_2+n_3) \to \kappa_i > 0$ for $\min\{n_1,n_2,n_3\} \to \infty$, $i \in \{1,2,3\}$.

... Several technical assumptions on the subject-effect functions

$$\mathbf{v}_j(t) = \mathbf{x}_j(t) - \boldsymbol{\eta}(t), j \in \{1,\dots,n\},\ t \in \mathcal{T}.$$

$\longrightarrow$ Assumed to be i.i.d. realisations of a stochastic process $\{\mathbf{V}(t) : t \in \mathcal{T}\}$, $\mathbf{V} : \Omega \to C^2(\mathcal{T})$ with zero mean function and covariance function $\mathbf{\Gamma}$.

$\longrightarrow$ Allow to use empirical process theory by van der Vaart and Wellner (2023).

Estimators: Pointwise Mean and Covariance

- Unbiased estimators for the mean function $\boldsymbol{\eta}$ is the point-wise mean:

$$\hat{\boldsymbol{\eta}}(t) = (\hat{\eta}_1(t), \hat{\eta}_2(t)) := \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j(t),\ t \in \mathcal{T}.$$

- Empirical covariance function $\hat{\boldsymbol{\Gamma}} : \mathcal{T}^2 \to \mathbb{R}^{2\times 2}$ at point (s,t):

$$\begin{aligned}\hat{\boldsymbol{\Gamma}}(s,t) &= \begin{pmatrix} \hat{\gamma}_1(s,t) & \hat{\gamma}_{12}(s,t) \\ \hat{\gamma}_{12}(s,t) & \hat{\gamma}_2(s,t) \end{pmatrix} \\ &:= \frac{1}{n-1}\sum_{j=1}^{n} \left(\mathbf{x}_j(s) - \hat{\boldsymbol{\eta}}(s)\right)\left(\mathbf{x}_j(t) - \hat{\boldsymbol{\eta}}(t)\right)',\ s,t \in \mathcal{T}.\end{aligned}$$

$\longrightarrow$ Point-wisely consistent for $\Gamma(s,t)$ by the weak law of large numbers for every $s,t \in \mathcal{T}$.
$\longrightarrow$ Uniformly consistent over $\{(t,t) \in \mathcal{T}^2 \mid t \in \mathcal{T}\}$ under the assumptions by Munko et al. (2025).

The Test's Construction: Two Point-wise Test Statistics

For the complete data part $\mathbf{x}^{(c)}$, consider a paired two-sample test statistic:

$$T_{n_1}\left(\mathbf{x}^{(c)}, \hat{\mathbf{\Gamma}}^{(c)}\right)(t) := \sqrt{n_1}\,\frac{\hat{\eta}_1^{(c)}(t) - \hat{\eta}_2^{(c)}(t)}{\sqrt{\hat{\gamma}_1^{(c)}(t,t) - 2\hat{\gamma}_{12}^{(c)}(t,t) + \hat{\gamma}_2^{(c)}(t,t)}},$$

and for the missing data part $\mathbf{x}^{(m)}$, as suggested in Amro and Pauly (2017), consider a point-wise Welch-type test statistic:

$$T_{n_2+n_3}\left(\mathbf{x}^{(m)}, \hat{\mathbf{\Gamma}}_0^{(m)}\right)(t) := \frac{\hat{\eta}_1^{(m)}(t) - \hat{\eta}_2^{(m)}(t)}{\sqrt{\frac{\hat{\gamma}_1^{(m)}(t,t)}{n_2} + \frac{\hat{\gamma}_2^{(m)}(t,t)}{n_3}}}, t \in \mathcal{T}.$$

The Test's Construction: Combination I

Weight the point-wise squared test statistics for every $t \in \mathcal{T}$:

$$K_n(\mathbf{x}, t) = \sqrt{a_n}\, T_{n_1}^2 \left(\mathbf{x}^{(c)}, \hat{\mathbf{\Gamma}}^{(c)} \right)(t) + \sqrt{1 - a_n} T_{n_2+n_3}^2 \left(\mathbf{x}^{(m)}, \hat{\mathbf{\Gamma}}_0^{(m)} \right)(t),$$

where $a_n \in [0, 1]$, $a_n \to a$.
Global test statistics:

$$I_n(\mathbf{x}) = \int_{\mathcal{T}} K_n(\mathbf{x}, t) dt,$$
$$S_n(\mathbf{x}) = \sup_{t \in \mathcal{T}} K_n(\mathbf{x}, t).$$

The Test's Construction: Combination II

Theorem

Consider the assumptions.

1. *Under $\mathcal{H}_0$, the test statistics $I(\mathbf{x})$ and $S(\mathbf{x})$ converge in distribution:*

$$I_n(\mathbf{x}) \xrightarrow{d} Z^I = \int_{\mathcal{T}} \sqrt{a} Z_1^2 + \sqrt{1-a} Z_2^2 dt, \quad S_n(\mathbf{x}) \xrightarrow{d} Z^S = \sup_{t \in \mathcal{T}} \sqrt{a} Z_1^2 + \sqrt{1-a} Z_2^2, \quad (2)$$

   where $Z_1 \sim \mathrm{GP}(0, \mathrm{Cov}(Z_1))$ and $Z_2 \sim \mathrm{GP}(0, \mathrm{Cov}(Z_2))$ with and $Var(Z_i) = 1$, $i \in \{1, 2\}$, for all $s, t \in \mathcal{T}$.

2. *Under $\mathcal{H}_1$, $I_n(\mathbf{x})$ and $S_n(\mathbf{x})$ converge in probability to ∞.*

The Test's Construction: Testing Procedures

$\longrightarrow$ Resampling Methods need to be considered.

- **Permutation Procedure**: uniform limiting distribution of the permutation test statistic $(T^{2,*}_{n_2+n_3}(t))_{t\in\mathcal{T}}$ differs from the limiting distribution of $(T^2_{n_2+n_3}(t))_{t\in\mathcal{T}}$.
  $\longrightarrow$ Equality holds, if $n_2 = n_3$ or $\gamma_1(s,t) = \gamma_2(s,t)$ for all $s,t \in \mathcal{T}$.
- **Parametric Bootstrap Procedure** is theoretically valid by Janssen and Pauls (2003) and Munko et al. (2025).
- Testing Procedures for some resampling method $\circ$:

$$\varphi_n^{\circ,I} = \mathbf{1}\left\{I_n(\mathbf{x}) > q_{1-\alpha}^{\circ,I}(\mathbf{x})\right\}; \quad \varphi_n^{\circ,S} = \mathbf{1}\left\{S_n(\mathbf{x}) > q_{1-\alpha}^{\circ,S}(\mathbf{x})\right\}.$$

Simulations

Simulation Set-Up

Consider the data generation model for $t \in [0,1]$, $j \in \{1,\ldots,n\}$ and $i \in \{1,2\}$:

$$x_{ij}(t) = c_i(t) + e_{ij}(t).$$

- Random errors $e_{ij}(t)$ based on standard Brownian Bridges and Zhang processes (Zhang, 2011) with different correlations.
  - Under the null:
  $$c_1(t) = c_2(t) = (\sin(2\pi t^2))^5 I_{[0,1]}(t).$$
  - Under the alternative:
  $$c_1(t) = (\sin(2\pi t^2))^5 I_{[0,1]}(t),\ c_2(t) = (\sin(2\pi t^2))^3 I_{[0,1]}(t).$$
- Small and moderate sample sizes (span $n = 15$ to $n = 60$).
- Missing values: randomly inserted by a MCAR mechanism until the samples n_2 and n_3 are achieved.
- Number of extracted points: $b \in \{25, 50\}$.

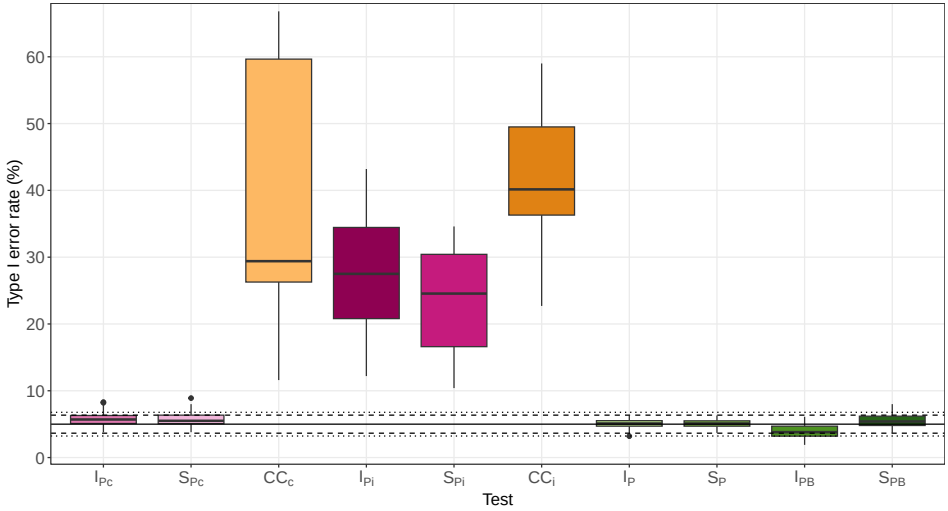
Simulation: Benchmark and New Methods

- I_{Pc} and S_{Pc}: permutation tests based on integral and supremum test statistics, by Smaga (2020) applied to the complete data $\mathbf{x}^{(\mathbf{c})}$.
- CC_c: bootstrap test by Crainiceanu et al. (2012) applied to complete data $\mathbf{x}^{(\mathbf{c})}$.

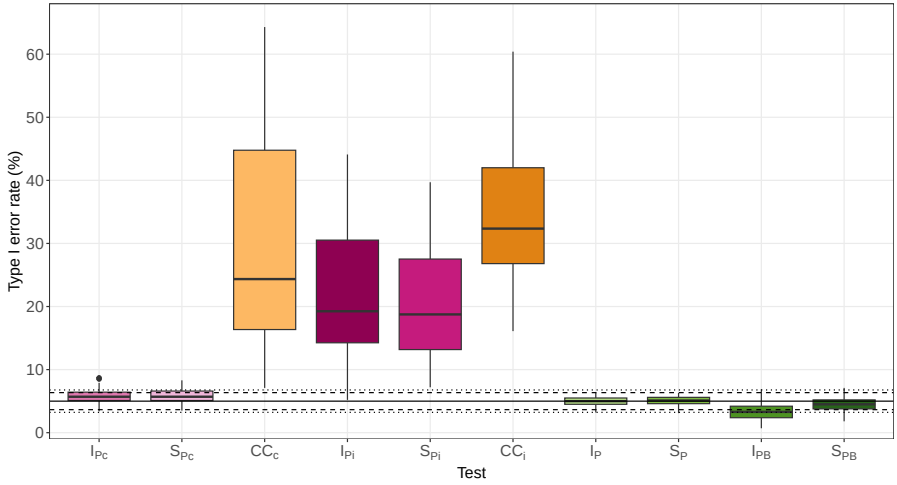
$\longrightarrow$ Do not use the whole data.

- I_{Pi}, S_{Pi}, CC_i: tests above applied to imputed data using the method by Jang et al. (2021).
- I_P and S_P: new permutation tests based on integral and supremum test statistics.
- I_{PB} and S_{PB}: new parametric bootstrap tests based on integral and supremum test statistics.

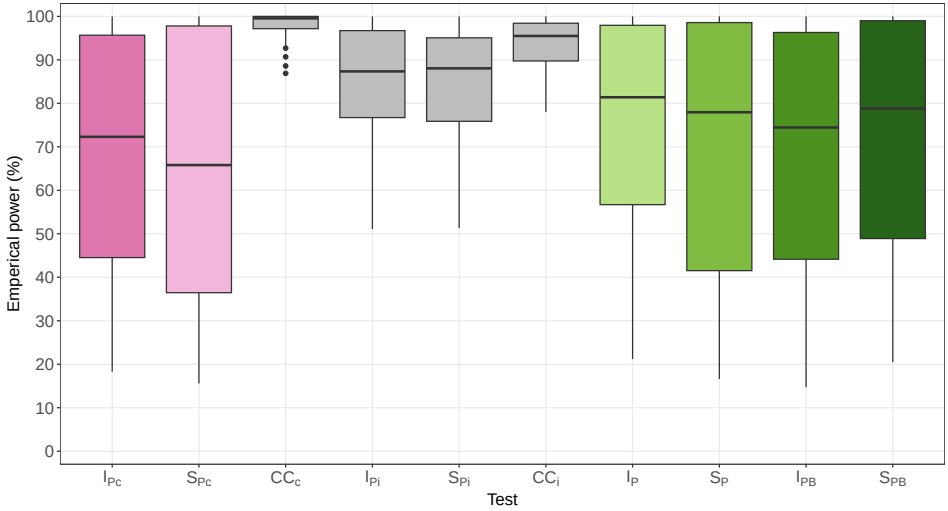
Simulation Results: Type I Error Brownian Bridges



Simulation Results: Type I Error Zhang Processes



Simulation Results: Power Brownian Bridges



Summary

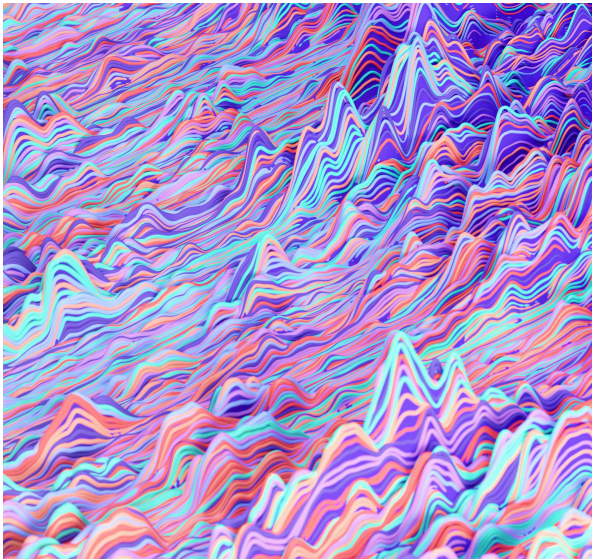
Summary

- Powerful method to consider missingness in functional data.
- Adaption of the permutation approach in Amro and Pauly (2017) is not theoretically valid, but looks numerically well in simulations.

$\longrightarrow$ We have proven the permutation method in Smaga (2020).

- Parametric bootstrap approach is theoretically valid, and performs similarly well in the simulations.
- Publication in preparation.

Thank you for your attention!



References I

- Amro, Lubna and Markus Pauly (2017). "Permuting Incomplete Paired Data: A Novel Exact and Asymptotic Correct Randomization Test". In: *Journal of Statistical Computation and Simulation* 87.6, pp. 1148–1159.
- Crainiceanu, Ciprian M. et al. (2012). "Bootstrap-Based Inference on the Difference in the Means of Two Correlated Functional Processes". In: *Statistics in Medicine* 31.26, pp. 3223–3240.
- Dette, Holger, Kevin Kokot, and Alexander Aue (2020). "Functional Data Analysis in the Banach Space of Continuous Functions". In: *The Annals of Statistics* 48.2, pp. 1168–1192.
- Jang, Jeong Hoon et al. (Sept. 30, 2021). "A Bayesian Multiple Imputation Approach to Bivariate Functional Data with Missing Components". In: *Statistics in medicine* 40.22, pp. 4772–4793.
- Janssen, Arnold and Thorsten Pauls (June 2003). "How Do Bootstrap and Permutation Tests Work?" In: *The Annals of Statistics* 31.3, pp. 768–806.

References II

- Martínez-Camblor, Pablo and Norberto Corral (Dec. 1, 2011). "Repeated Measures Analysis for Functional Data". In: *Computational Statistics & Data Analysis* 55.12, pp. 3244–3256.
- Munko, Merle et al. (Jan. 2025). "Multiple Comparison Procedures for Simultaneous Inference in Functional MANOVA". In: *Electronic Journal of Statistics* 19.1, pp. 2637–2732.
- Smaga, Łukasz (Mar. 1, 2020). "A Note on Repeated Measures Analysis for Functional Data". In: *AStA Advances in Statistical Analysis* 104.1, pp. 117–139.
- Van der Vaart, A. W. and Jon A. Wellner (2023). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Cham: Springer International Publishing.
- Zhang, Jin-Ting (2011). "Statistical Inferences for Linear Models with Functional Responses". In: *Statistica Sinica* 21.3, pp. 1431–1451.

Assumptions

1. The subject-effect functions $\mathbf{v}_j(t) = \mathbf{x}_j(t) - \boldsymbol{\eta}(t)$, $j \in \{1, \ldots, n\}$, $t \in \mathcal{T}$, are i.i.d. realisations of a stochastic process $\{\mathbf{V}(t) : t \in \mathcal{T}\}$, $\mathbf{V} : \Omega \to C^2(\mathcal{T})$ with zero mean function and covariance function $\boldsymbol{\Gamma}$.
2. There exist $t^* \in \mathcal{T}$ and $J \leq 2$ such that $\mathrm{E}[\|\mathbf{v}_1(t^*)\|_2^J] < \infty$, where here and throughout $\|\cdot\|_2$ denotes the Euclidean norm of a vector of numbers.
3. There exists a $C \geq 0$ for all $s, t \in \mathcal{T}$ such that $\mathrm{E}[\|\mathbf{v}_1(t) - \mathbf{v}_1(s)\|_2^J] \leq C \cdot \rho(s,t)^J$, and there is a $\nu > 0$ such that

$$\int_0^\nu \mathrm{D}(\epsilon, \rho)^{2/J} d\epsilon < \infty. \tag{3}$$

Here, and throughout, J is chosen as in (2) and $\mathrm{D}(\epsilon, \rho)$ denotes the packing number, see van der Vaart and Wellner (2023, Def. 2.2.3).

Covariances of the Limiting Processes of the Paired and Unpaired Test Statistics

$\mathrm{Var}(Z_i) = 1$, $i \in \{1, 2\}$, for all $s, t \in \mathcal{T}$ and:

- $Z_1 \sim \mathrm{GP}(0, \mathrm{Cov}(Z_1))$ with

$$\mathrm{Cov}(Z_1(t), Z_1(s)) = \frac{\gamma_{1-2}(s,t)}{\sqrt{\gamma_{1-2}(s)\gamma_{1-2}(t)}}.$$

- $Z_2 \sim \mathrm{GP}(0, \mathrm{Cov}(Z_2))$ with

$$\mathrm{Cov}(Z_2(t), Z_2(s)) = \frac{\frac{1}{\kappa_2}\gamma_1(s,t) + \frac{1}{\kappa_3}\gamma_2(s,t)}{\sqrt{\frac{1}{\kappa_2}\gamma_1(t,t) + \frac{1}{\kappa_3}\gamma_2(t,t)}\sqrt{\frac{1}{\kappa_2}\gamma_1(s,s) + \frac{1}{\kappa_3}\gamma_2(s,s)}}.$$