Reconstructing partially observed functional data via factor models of increasing rank

Maximilian Ofner Siegfried Hörmann

Graz University of Technology
Institute of Statistics

COMPSTAT 2023
London

Outline

1. Introduction
2. Estimation of reconstructions
3. Simultaneous prediction bands
4. Real data illustration

Section 1

Introduction

Let

$$X = (X(u) : u \in [0,1])$$

be a (centred) random function *observable* on $O \subset [0,1]$.

Let

$$X = (X(u) : u \in [0, 1])$$

be a (centred) random function *observable* on $\mathrm{O} \subset [0, 1]$.

Assumption (MCAR)

The set O is independent of X.

See Liebl and Rameseder (2019) for a relaxation of this assumption.

Let

$$X = (X(u) : u \in [0,1])$$

be a (centred) random function *observable* on $O \subset [0,1]$.

Assumption (MCAR)

The set O is independent of X.

See Liebl and Rameseder (2019) for a relaxation of this assumption.

Assume $X^O = (X(u) : u \in O)$ satisfies the Karhunen-Loève expansion

$$X^O(u) = \sum_{k=1}^{\infty} \xi_k^O \phi_k^O(u), \quad u \in O,$$

where

- (ξ_k^O) are uncorrelated rv's with $\mathbb{V}\mathrm{ar}(\xi_k^O) = \lambda_k^O$,
- (ϕ_k^O) are ONB of $L^2(O)$.

Assume $\lambda_1^O \geq \lambda_2^O \geq \ldots$.

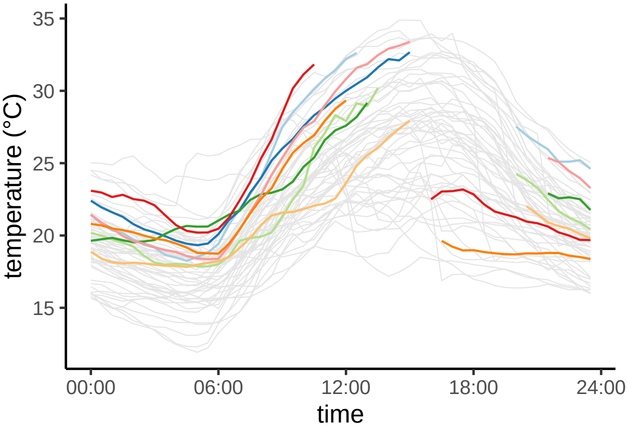


Figure: Intraday temperature values measured half-hourly in the east of Graz (Austria) between 1 July and 31 August 2022. Complete observations in grey.

Data: Land Steiermark (2023)

Question

How can we infer X from X^O?

Question

How can we infer X from X^{O}?

Kneip and Liebl (2020) study the linear operator $\mathcal{L} : L^2(O) \to L^2[0,1]$,

$$\mathcal{L}(X^{O})(u) = \sum_{k=1}^{\infty} \xi_k^{O} \tilde{\phi}_k^{O}(u), \qquad u \in [0,1],$$

where

$$\tilde{\phi}_k^{O}(u) = \frac{\mathbb{E}\left[X(u)\xi_k^{O}\right]}{\lambda_k^{O}}, \qquad u \in [0,1],$$

are *extrapolated basis functions*.

Question

How can we infer X from X^O?

Kneip and Liebl (2020) study the linear operator $\mathcal{L} : L^2(O) \to L^2[0,1]$,

$$\mathcal{L}(X^O)(u) = \sum_{k=1}^{\infty} \xi_k^O \tilde{\phi}_k^O(u), \qquad u \in [0,1],$$

where

$$\tilde{\phi}_k^O(u) = \frac{\mathbb{E}\left[X(u)\xi_k^O\right]}{\lambda_k^O}, \qquad u \in [0,1],$$

are *extrapolated basis functions*. It holds

$$X(u) = \mathcal{L}(X^O)(u) + Z^O(u), \qquad u \in [0,1],$$

for some *reconstruction error* $Z^O = (Z^O(u) : u \in [0,1])$.

We want to avoid restrictive smoothness conditions such as differentiable paths.

Idea

Estimate $\mathcal{L}(X^{O})$ via approximate factor models.

Assumption (Rank)

There exists some $r_{O} < \infty$ such that

$$X^{O}(u) = \sum_{k=1}^{r_{O}} \xi_k^{O} \varphi_k^{O}(u), \qquad u \in O.$$

Later: $r_{O} \to \infty$ to account for high dimensionality.

Let $\{X_t : t \leq T\}$ be iid copies of X. Consider a regular grid (size N)

$$0 = u_1 < u_2 < \cdots < u_N = 1.$$

Let $\{X_t : t \leq T\}$ be iid copies of X. Consider a regular grid (size N)

$$0 = u_1 < u_2 < \cdots < u_N = 1.$$

For e_{ti} independent of $\{X_t : t \leq T\}$, not necessarily iid, we observe

$$Y_{ti} = X_t(u_i) + e_{ti},$$

whenever $u_i \in \mathrm{O}_t \supset \mathrm{O}$.

Let $\{X_t : t \leq T\}$ be iid copies of X. Consider a regular grid (size N)

$$0 = u_1 < u_2 < \cdots < u_N = 1.$$

For e_{ti} independent of $\{X_t : t \leq T\}$, not necessarily iid, we observe

$$Y_{ti} = X_t(u_i) + e_{ti},$$

whenever $u_i \in \mathrm{O}_t \supset \mathrm{O}$. For $u_i \in \mathrm{O}$, the rank assumption implies

$$Y_{ti} = \sum_{k=1}^{r_\mathrm{O}} \xi_{tk}^\mathrm{O} \phi_k^\mathrm{O}(u_i) + e_{ti} = F_t^\mathrm{O} \Lambda_i^\mathrm{O} + e_{ti},$$

where we define

- $F_t^\mathrm{O} = \left(\frac{\xi_{t1}^\mathrm{O}}{\sqrt{\lambda_1^\mathrm{O}}}, \ldots, \frac{\xi_{tr_\mathrm{O}}^\mathrm{O}}{\sqrt{\lambda_{r_\mathrm{O}}^\mathrm{O}}} \right)$,
- $\Lambda_i^\mathrm{O} = \left(\sqrt{\lambda_1^\mathrm{O}} \phi_1^\mathrm{O}(u_i), \ldots, \sqrt{\lambda_{r_\mathrm{O}}^\mathrm{O}} \phi_{r_\mathrm{O}}^\mathrm{O}(u_i) \right)'$.

Approximate factor model

If the e_{ti} are only mildly correlated, then

$$Y_{ti} = F_t^{\circ} \Lambda_i^{\circ} + e_{ti},$$

constitutes an *approximate factor model* (AFM) of rank $r_{\circ}$.

A major branch of works analyses AFMs under a double asymptotic

$$N, T \to \infty.$$

Important contributions include Chamberlain and Rothschild (1983), Bai and Ng (2002), Bai (2003), and Fan et al. (2013). Imputation of missing values is studied by Bai and Ng (2021), Cahan et al. (2023), and Xiong and Pelger (2023).

Important features of our factor model

- The rank r_O depends on $O \subset [0, 1]$.
- We allow the rank r_O to grow and consider a triple asymptotic

$$r_O, N, T \to \infty.$$

  See Li et al. (2017) and Hörmann and Jammoul (2022) for a similar asymptotic.

- The eigenvalue λ_k^O measures the pervasiveness of the k-th factor score and enters our convergence rates.

  See Bai and Ng (2023) for related results on AFMs with weaker loadings.

Section 2

Estimation of reconstructions

Estimation

1. Let $\widehat{F_t^{\mathrm{O}}}$ $(1 \times r_{\mathrm{O}})$ be the PC-estimate of the t-th factors.
2. Assume the first $T_{\mathrm{C}} \leq T$ curves are completely observable.
   - $\widehat{\boldsymbol{F}^{\mathrm{O}}}$ ... $(T_{\mathrm{C}} \times r_{\mathrm{O}})$ matrix with $\widehat{F_t^{\mathrm{O}}}$ in its rows, $t \leq T_{\mathrm{C}}$.
   - Y_i ... $(1 \times T_{\mathrm{C}})$ vector with elements Y_{ti}, $t \leq T_{\mathrm{C}}$.

Convention: $\frac{\widehat{\boldsymbol{F}^{\mathrm{O}}}'\widehat{\boldsymbol{F}^{\mathrm{O}}}}{T_{\mathrm{C}}} = \boldsymbol{I}$.

3. Estimate $\mathcal{L}(X_t^{\mathrm{O}})(u_i)$ by projecting Y_i onto the estimated factors,

$$\widehat{\mathcal{L}(X_t^{\mathrm{O}})}(u_i) = \widehat{F_t^{\mathrm{O}}}(\widehat{\boldsymbol{F}^{\mathrm{O}}}'\widehat{\boldsymbol{F}^{\mathrm{O}}})^{-1}\widehat{\boldsymbol{F}^{\mathrm{O}}}'Y_i = \frac{\widehat{F_t^{\mathrm{O}}}\widehat{\boldsymbol{F}^{\mathrm{O}}}'Y_i}{T_c}$$

and interpolate linearly.

Theorem (OH, 2023)

Let X be observable on $\mathrm{O} \subset [0,1]$. Under regularity conditions,

$$\sup_{u \in [0,1]} |\widehat{\mathcal{L}(X^{\mathrm{O}})}(u) - \mathcal{L}(X^{\mathrm{O}})(u)| = O_p\left(\frac{r_{\mathrm{O}}}{\lambda^{\mathrm{O}}_{r_{\mathrm{O}}}} \sqrt{\frac{1}{N_{\mathrm{O}}} + \frac{\log(N)}{T_{\mathrm{C}}}} \right),$$

as $r_{\mathrm{O}} \to \infty$, $N_{\mathrm{O}} \to \infty$ with $N \to \infty$ and $T_{\mathrm{C}} \to \infty$ with $T \to \infty$.

r_{O}	...	rank of X^{O} on O,
N_{O}	...	number of grid points $u_i \in \mathrm{O}$,
T_{C}	...	number of complete curves.

If $\mathrm{O} = [0,1]$, then $\mathcal{L}(X^{\mathrm{O}}) = X$ and the estimator $\widehat{X} = \widehat{\mathcal{L}(X^{\mathrm{O}})}$ can be seen as a smoother of X.

Theorem (OH, 2023)

Let X be observable on $O \subset [0,1]$. *Under regularity conditions,*

$$\sup_{u \in [0,1]} |\widehat{\mathcal{L}(X^O)}(u) - \mathcal{L}(X^O)(u)| = O_p\left(\frac{r_O}{\lambda^O_{r_O}}\sqrt{\frac{1}{N_O} + \frac{\log(N)}{T_C}}\right),$$

as $r_O \to \infty$, $N_O \to \infty$ *with* $N \to \infty$ *and* $T_C \to \infty$ *with* $T \to \infty$.

r_O	...	rank of X^O on O,
N_O	...	number of grid points $u_i \in O$,
T_C	...	number of complete curves.

If $O = [0,1]$, then $\mathcal{L}(X^O) = X$ and the estimator $\widehat{X} = \widehat{\mathcal{L}(X^O)}$ can be seen as a smoother of X.

Theorem (OH, 2023)

Let X be observable on $\mathrm{O} \subset [0,1]$. *Under regularity conditions,*

$$\sup_{u\in[0,1]} |\widehat{\mathcal{L}(X^{\mathrm{O}})}(u) - \mathcal{L}(X^{\mathrm{O}})(u)| = O_p\left(\frac{r_{\mathrm{O}}}{\lambda^{\mathrm{O}}_{r_{\mathrm{O}}}}\sqrt{\frac{1}{N_{\mathrm{O}}} + \frac{\log(N)}{T_{\mathrm{C}}}}\right),$$

as $r_{\mathrm{O}} \to \infty$, $N_{\mathrm{O}} \to \infty$ *with* $N \to \infty$ *and* $T_{\mathrm{C}} \to \infty$ *with* $T \to \infty$.

r_{O}	...	rank of X^{O} on O,
N_{O}	...	number of grid points $u_i \in \mathrm{O}$,
T_{C}	...	number of complete curves.

If $\mathrm{O} = [0,1]$, then $\mathcal{L}(X^{\mathrm{O}}) = X$ and the estimator $\widehat{X} = \widehat{\mathcal{L}(X^{\mathrm{O}})}$ can be seen as a smoother of X.

Theorem (OH, 2023)

Let X be observable on $O \subset [0,1]$. Under regularity conditions,

$$\sup_{u \in [0,1]} |\widehat{\mathcal{L}(X^O)}(u) - \mathcal{L}(X^O)(u)| = O_p\left(\frac{r_O}{\lambda^O_{r_O}} \sqrt{\frac{1}{N_O} + \frac{\log(N)}{T_C}} \right),$$

as $r_O \to \infty$, $N_O \to \infty$ with $N \to \infty$ and $T_C \to \infty$ with $T \to \infty$.

r_O	...	rank of X^O on O,
N_O	...	number of grid points $u_i \in O$,
T_C	...	number of complete curves.

If $O = [0,1]$, then $\mathcal{L}(X^O) = X$ and the estimator $\widehat{X} = \widehat{\mathcal{L}(X^O)}$ can be seen as a smoother of X.

Section 3

Simultaneous prediction bands

Question

How can we construct simultaneous prediction bands (SPB) for the reconstructions?

Assume $\widehat{X}_t$ and $\widehat{\mathcal{L}(X_t^{\mathrm{O}})}$ consistently estimate X_t and $\mathcal{L}(X_t^{\mathrm{O}})$. Set

$$\widehat{Z}_t^{\mathrm{O}}(u) = \begin{cases} \widehat{X}_t(u) - \widehat{\mathcal{L}(X_t^{\mathrm{O}})}(u), & u \in \mathrm{M} = [0,1] \setminus \mathrm{O}, \\ 0, & u \in \mathrm{O}, \end{cases}$$

and let $\widehat{q}_\alpha$ be an estimator of q_α defined by

$$\mathbb{P}\big(\sup_{u \in \mathrm{M}} \{|Z^{\mathrm{O}}(u)|/\mathrm{sd}(Z^{\mathrm{O}}(u))\} > q_\alpha\big) = \alpha.$$

For $\alpha \in (0,1)$, we then consider the SPB

$$\widehat{\mathcal{L}(X_t^{\mathrm{O}})}(u) \pm \widehat{q}_\alpha \widehat{\mathrm{sd}}(Z^{\mathrm{O}}(u)), \qquad u \in \mathrm{M}.$$

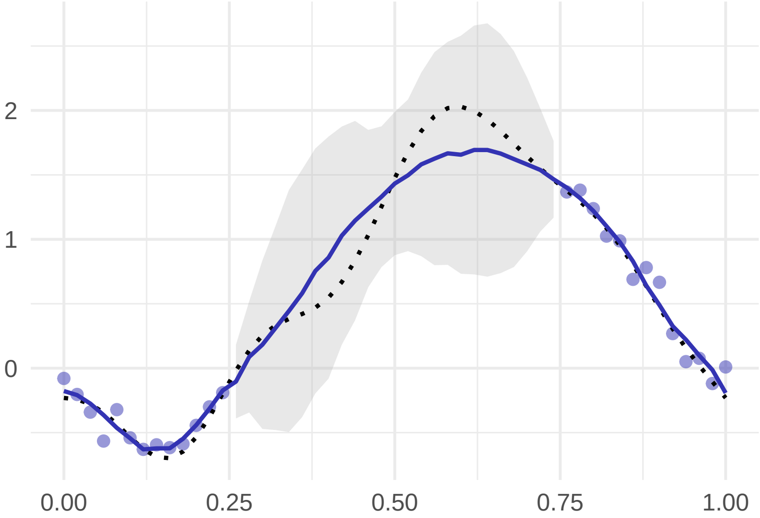


Figure: Reconstruction of a simulated curve along with 95% prediction band.

Section 4

Real data illustration

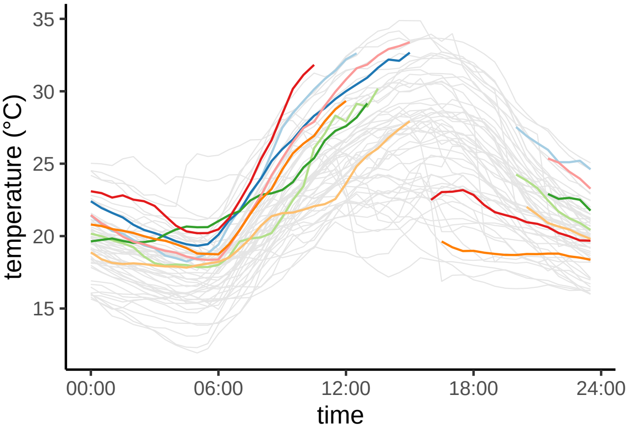
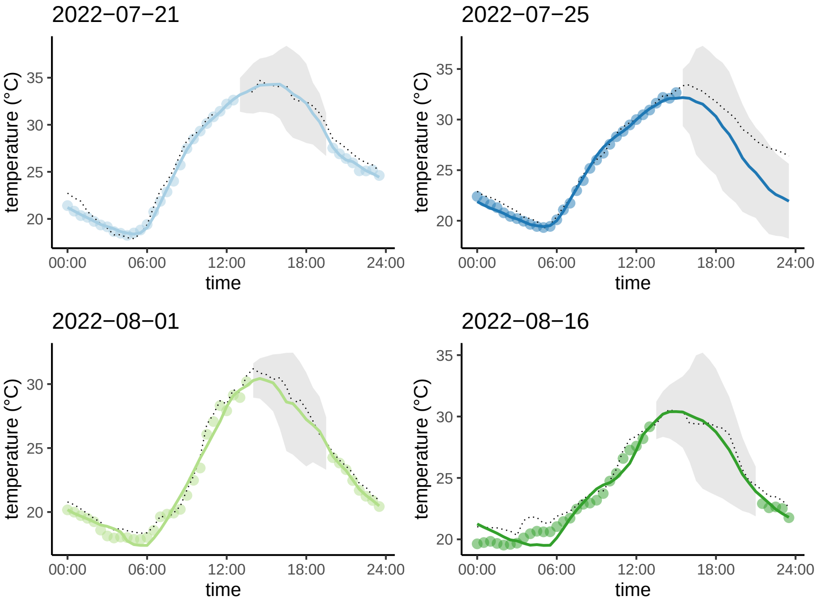
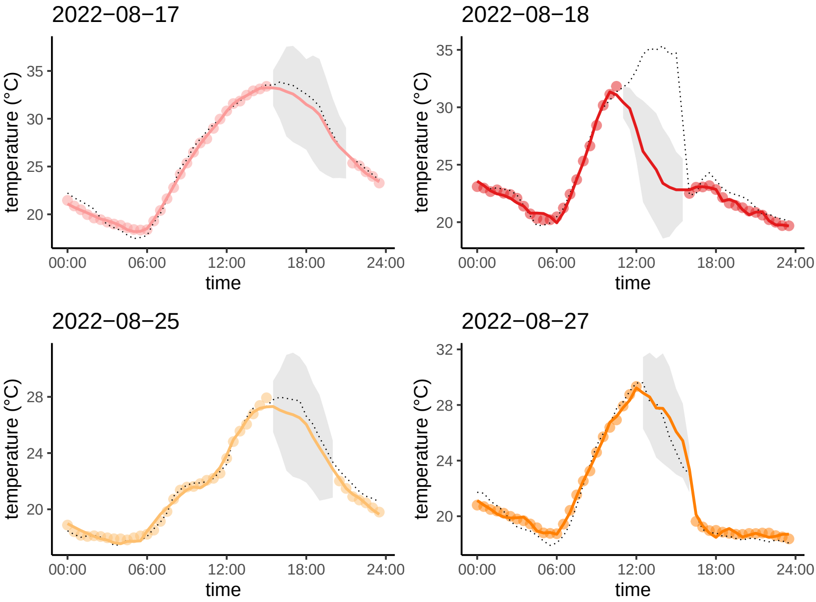


Figure: Intraday temperature values measured half-hourly in the east of Graz (Austria) between 1 July and 31 August 2022. Complete observations in grey.


2022−07−21
2022−07−25
2022−08−01
2022−08−16
temperature (°C)
time
00:00
06:00
12:00
18:00
24:00
20
25
30
35


2022−08−17
2022−08−18
2022−08−25
2022−08−27
temperature (°C)
time
00:00
06:00
12:00
18:00
24:00
20
25
30
35
24
28
32

Questions? Comments? Criticism?

Manuscript: arXiv:2305.13152
Code: Github: FDFM

Maximilian Ofner
m.ofner@tugraz.at

Graz University of Technology

References

Bai, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, 71(1):135–171.

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1):191–221.

Bai, J. and Ng, S. (2021). Matrix completion, counterfactuals, and factor analysis of missing data. *J. Amer. Statist. Assoc.*, 116(536):1746–1763.

Bai, J. and Ng, S. (2023). Approximate factor models with weaker loadings. *Journal of Econometrics*.

Cahan, E., Bai, J., and Ng, S. (2023). Factor-based imputation of missing values and covariances in panel data of large dimensions. *J. Econometrics*, 233(1):113–131.

Chamberlain, G. and Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5):1281–1304.

Fan, J., Liao, Y., and Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 75(4):603–680.

Hörmann, S. and Jammoul, F. (2022). Consistently recovering the signal from noisy functional data. *J. Multivariate Anal.*, 189:Paper No. 104886, 18.

Kneip, A. and Liebl, D. (2020). On the optimal reconstruction of partially observed functional data. *Ann. Statist.*, 48(3):1692–1717.

Land Steiermark (2023). Air quality data. `https://app.luis.steiermark.at/luft2/suche.php`,. Licensed under CC BY 4.0; accessed on March 28, 2023.

Li, H., Li, Q., and Shi, Y. (2017). Determining the number of factors when the number of factors can increase with sample size. *J. Econometrics*, 197(1):76–86.

Liebl, D. and Rameseder, S. (2019). Partially observed functional data: the case of systematically missing parts. *Comput. Statist. Data Anal.*, 131:104–115.

Xiong, R. and Pelger, M. (2023). Large dimensional latent factor modeling with missing observations and applications to causal inference. *J. Econometrics*, 233(1):271–301.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.*, 100(470):577–590.