Modelling Function-Valued Processes with Nonseparable and/or Nonstationary Covariance Structure

Jian Qing Shi

School of Mathematics, Statistics & Physics
Newcastle University, and
Turing Fellow, Alan Turing Institute, UK

Joint work with Evandro Konzen (Reading, UK) and Zhanfeng Wang (USTC)

International Statistical Conference in
Memory of Professor Sik-Yum Lee, 17-18/12/2019, CUHK, HK

International Statistical Conference in Memory of Prof. S Y Lee

'When you identify the problems, you finish half of the project.'

Overview

1. Multi-dimensional function-valued processes
   - Covariance separability assumption
2. Bayesian process regression analysis
   - Stationary model
   - Nonstationary GPs
3. Numerical results
4. Conclusions

Multi-dimensional function-valued processes

In FPCA, the random process $X(\boldsymbol{t})$, $\boldsymbol{t} \in \mathbb{R}^Q$, is represented as (Karhunen-Loéve expansion)

$$X(\boldsymbol{t}) = \mu(\boldsymbol{t}) + \sum_{j=1}^{\infty} \xi_j \nu_j(\boldsymbol{t}),$$

where ξ_j are uncorrelated random variables and ν_j are eigenfunctions of the covariance operator of X, i.e. ν_j are solutions to the equation

$$\int k(\boldsymbol{t}, \boldsymbol{t}')\nu(\boldsymbol{t}')d\boldsymbol{t}' = \lambda\nu(\boldsymbol{t}).$$

The eigenvalue λ_j is the variance of X in the principal direction ν_j and the cumulative fraction of variance explained by the first J directions is given by

$$\mathrm{CFVE}_J = \frac{\sum_{j=1}^{J} \lambda_j}{\sum_{j=1}^{M} \lambda_j}, \quad \text{where } M \text{ is large.}$$

Multi-dimensional function-valued processes

In FPCA, the random process $X(\boldsymbol{t}), \boldsymbol{t} \in \mathbb{R}^Q$, is represented as (Karhunen-Loéve expansion)

$$X(\boldsymbol{t}) = \mu(\boldsymbol{t}) + \sum_{j=1}^{\infty} \xi_j \nu_j(\boldsymbol{t}),$$

where ξ_j are uncorrelated random variables and ν_j are eigenfunctions of the covariance operator of X, i.e. ν_j are solutions to the equation

$$\int k(\boldsymbol{t}, \boldsymbol{t}')\nu(\boldsymbol{t}')d\boldsymbol{t}' = \lambda\nu(\boldsymbol{t}).$$

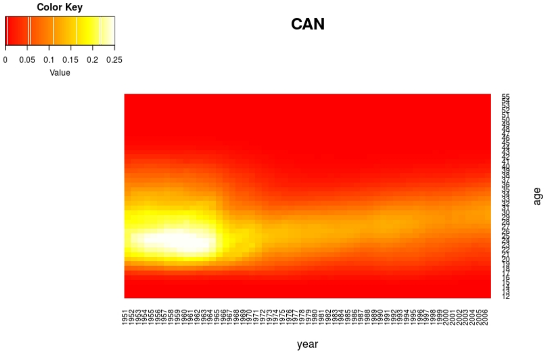
The eigenvalue λ_j is the variance of X in the principal direction ν_j and the cumulative fraction of variance explained by the first J directions is given by

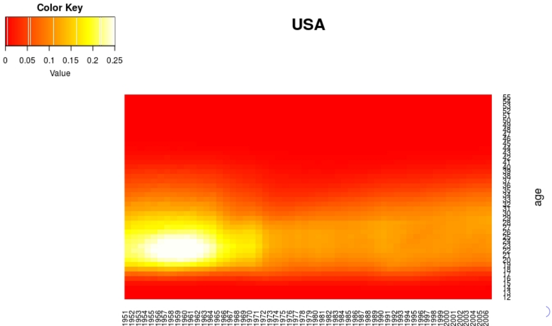
$$\text{CFVE}_J = \frac{\sum_{j=1}^{J} \lambda_j}{\sum_{j=1}^{M} \lambda_j}, \quad \text{where } M \text{ is large.}$$

When $Q = 1$, the method is well developed; but it is challenging when Q is large.

Human Fertility Data

- Age-Specific Fertility Rate (ASFR) for country j: $X_j(s,t)$, $j = 1, \ldots, N$, $s \in \mathcal{S}, \quad t \in \mathcal{T}$.
- Observed data:
  - women's age: $s = 12, 13, \ldots, 55$
  - calendar year: $t = 1951, 1952, \ldots, 2006$





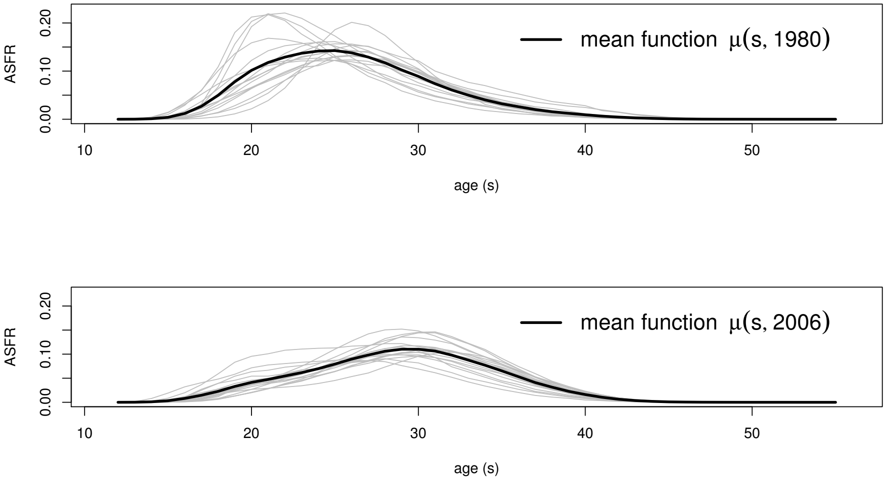


Figure 1: Human fertility rates of 17 countries over age for two different years.

Covariance function

We need to estimate

$$\mathrm{Cov}\Big(X(s,t), X(s',t')\Big) = k(s,t;s',t'),$$

Chen et al. (*JRSSb* 2017) suggest to use tensor product representations:

$$\text{Marginal FPCA:} \qquad X(s,t) = \mu(s,t) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \chi_{jk}\phi_{jk}(t)\psi_j(s)$$

$$\text{Product FPCA:} \qquad X(s,t) = \mu(s,t) + \sum_{j=1}^{\infty}\sum_{k=1}^{\infty} \chi_{jk}\phi_{k}(t)\psi_j(s)$$

For the Product FPCA, this means

$$\begin{aligned} k(s,t;s',t') &= \lim_{J\to\infty}\sum_{j=1}^{J}\sum_{k=1}^{J} \lambda_k\gamma_j\phi_k(t)\psi_j(s)\phi_k(t')\psi_j(s') \\ &= k_1(s,s')k_2(t,t'). \end{aligned}$$

Separability assumption of covariance functions

The covariance function of the random process $X(\boldsymbol{t}),\ \boldsymbol{t} \in \Re^2$, is said to be *separable* when

$$k(t_1, t_2; t_1', t_2') = k_1(t_1, t_1')k_2(t_2, t_2').$$

Main advantages:

- it reduces computational costs;
- it is easier to guarantee that the full covariance function is positive semi-definite.
- Covariance function for each coordinate can be estimated nonparametrically

Separability assumption of covariance functions

The covariance function of the random process $X(\boldsymbol{t})$, $\boldsymbol{t} \in \Re^2$, is said to be *separable* when

$$k(t_1, t_2; t_1', t_2') = k_1(t_1, t_1')k_2(t_2, t_2').$$

Main advantages:

- it reduces computational costs;
- it is easier to guarantee that the full covariance function is positive semi-definite.
- Covariance function for each coordinate can be estimated nonparametrically

Disadvantage:

- no interaction between t_1 and t_2 in the covariance structure is allowed.

Here we are not interested in interactions in the mean function:

$$E\big(X(\boldsymbol{t})\big) = \gamma_0 + \gamma_1(t_1) + \gamma_2(t_2) + \gamma_{12}(t_1, t_2)$$

Separability assumption of covariance functions

The covariance function of the random process $X(\boldsymbol{t})$, $\boldsymbol{t} \in \Re^2$, is said to be *separable* when

$$k(t_1, t_2; t_1', t_2') = k_1(t_1, t_1')k_2(t_2, t_2').$$

Main advantages:

- it reduces computational costs;
- it is easier to guarantee that the full covariance function is positive semi-definite.
- Covariance function for each coordinate can be estimated nonparametrically

Disadvantage:

- no interaction between t_1 and t_2 in the covariance structure is allowed.

Here we are not interested in interactions in the mean function:

$$E\big(X(\boldsymbol{t})\big) = \gamma_0 + \gamma_1(t_1) + \gamma_2(t_2) + \gamma_{12}(t_1, t_2)$$

Covariance separability implies separability of eigenfunctions.

Process regression model

$$X(\boldsymbol{t}) = \mu(\boldsymbol{t}) + f(\boldsymbol{t}) + \epsilon(\boldsymbol{t}), \quad f(\boldsymbol{t}), \ \boldsymbol{t} \in \Re^Q.$$

- To address the difficulties in the estimation of $k(\boldsymbol{t}, \boldsymbol{t}')$, we can model the random process f by a process prior.

Process regression model

$$X(\boldsymbol{t}) = \mu(\boldsymbol{t}) + f(\boldsymbol{t}) + \epsilon(\boldsymbol{t}), \quad f(\boldsymbol{t}), \; \boldsymbol{t} \in \Re^Q.$$

- To address the difficulties in the estimation of $k(\boldsymbol{t}, \boldsymbol{t}')$, we can model the random process f by a process prior.
- A Gaussian process regression (GPR) model (O'Hagan and Kingman, 1978; Rasmussen and Williams, 2006; Shi and Choi, 2011) is defined as:
  - the prior of $f(\boldsymbol{t})$ is a GP with zero mean, and
  - a covariance function

$$k(\cdot,\cdot) : \mathcal{T}^2 \to \mathbb{R}, \; k(\boldsymbol{t}, \boldsymbol{t}') = \mathrm{Cov}\Big[f(\boldsymbol{t}), f(\boldsymbol{t}')\Big].$$

Process regression model

$$X(\boldsymbol{t}) = \mu(\boldsymbol{t}) + f(\boldsymbol{t}) + \epsilon(\boldsymbol{t}), \quad f(\boldsymbol{t}), \ \boldsymbol{t} \in \Re^Q.$$

- To address the difficulties in the estimation of $k(\boldsymbol{t}, \boldsymbol{t}')$, we can model the random process f by a process prior.
- A Gaussian process regression (GPR) model (O'Hagan and Kingman, 1978; Rasmussen and Williams, 2006; Shi and Choi, 2011) is defined as:
  - the prior of $f(\boldsymbol{t})$ is a GP with zero mean, and
  - a covariance function

$$k(\cdot,\cdot) : \mathcal{T}^2 \to \mathbb{R}, \ k(\boldsymbol{t}, \boldsymbol{t}') = \mathrm{Cov}\Big[f(\boldsymbol{t}), f(\boldsymbol{t}')\Big].$$

  - Marginally, for any finite n and $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n \in \mathcal{T}$, the joint distribution of $X_n = \Big(X(\boldsymbol{t}_1), \ldots, X(\boldsymbol{t}_n)\Big)'$, if $\epsilon(\boldsymbol{t})$ is normal, is an n-variate Gaussian distribution with mean vector $\boldsymbol{\mu}_n = \Big(\mu(\boldsymbol{t}_1), \ldots, \mu(\boldsymbol{t}_n)\Big)'$ and covariance matrix Ψ_n whose (i,j)-th entry is given by $\Big[\Psi_n\Big]_{ij} = k(\boldsymbol{t}_i, \boldsymbol{t}_j) + \delta_{ij}\sigma_\varepsilon^2, \ i,j = 1, \ldots, n.$

Process regression model

$$X(\boldsymbol{t}) = \mu(\boldsymbol{t}) + f(\boldsymbol{t}) + \epsilon(\boldsymbol{t}), \quad f(\boldsymbol{t}), \ \boldsymbol{t} \in \Re^Q.$$

- To address the difficulties in the estimation of $k(\boldsymbol{t}, \boldsymbol{t}')$, we can model the random process f by a process prior.
- A Gaussian process regression (GPR) model (O'Hagan and Kingman, 1978; Rasmussen and Williams, 2006; Shi and Choi, 2011) is defined as:
  - the prior of $f(\boldsymbol{t})$ is a GP with zero mean, and
  - a covariance function

$$k(\cdot,\cdot) : \mathcal{T}^2 \to \mathbb{R}, \ k(\boldsymbol{t}, \boldsymbol{t}') = \mathrm{Cov}\Big[f(\boldsymbol{t}), f(\boldsymbol{t}')\Big].$$

  - Marginally, for any finite n and $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_n \in \mathcal{T}$, the joint distribution of $X_n = \Big(X(\boldsymbol{t}_1), \ldots, X(\boldsymbol{t}_n)\Big)'$, if $\epsilon(\boldsymbol{t})$ is normal, is an n-variate Gaussian distribution with mean vector $\boldsymbol{\mu}_n = \Big(\mu(\boldsymbol{t}_1), \ldots, \mu(\boldsymbol{t}_n)\Big)'$ and covariance matrix Ψ_n whose (i,j)-th entry is given by $\Big[\Psi_n\Big]_{ij} = k(\boldsymbol{t}_i, \boldsymbol{t}_j) + \delta_{ij}\sigma_\varepsilon^2, \ i,j = 1, \ldots, n.$
- If $\epsilon(\boldsymbol{t})$ or $X(\boldsymbol{t})$ is non-Gaussian, the marginal distribution is much more complicated (see e.g. Wang and Shi, 2014)

Parametric isotropic covariance functions

Powered Exponential:

$$k(\boldsymbol{t}, \boldsymbol{t}') = \nu \exp\Big\{ -\omega ||\boldsymbol{t} - \boldsymbol{t}'||^{\gamma} \Big\}, \quad \nu > 0, \quad \omega \geq 0, \quad 0 < \gamma \leq 2.$$

Rational Quadratic:

$$k(\boldsymbol{t}, \boldsymbol{t}') = \Big(1 + s_{\alpha} \omega ||\boldsymbol{t} - \boldsymbol{t}'||^2 \Big)^{-\alpha}, \quad \alpha, \omega \geq 0.$$

Matérn:

$$k(\boldsymbol{t}, \boldsymbol{t}') = \frac{1}{\Gamma(\nu) 2^{\nu-1}} \Big(\sqrt{2\nu} \omega ||\boldsymbol{t} - \boldsymbol{t}'|| \Big)^{\nu} \mathcal{K}_{\nu} \Big(\sqrt{2\nu} \omega ||\boldsymbol{t} - \boldsymbol{t}'|| \Big), \quad \omega \geq 0,$$

where $\mathcal{K}_{\nu}$ is the modified Bessel function of order ν.

Parametric isotropic covariance functions

Powered Exponential:

$$k(\boldsymbol{t}, \boldsymbol{t}') = \nu \exp\Big\{ -\omega ||\boldsymbol{t} - \boldsymbol{t}'||^{\gamma} \Big\}, \quad \nu > 0, \quad \omega \geq 0, \quad 0 < \gamma \leq 2.$$

Rational Quadratic:

$$k(\boldsymbol{t}, \boldsymbol{t}') = \Big(1 + s_{\alpha} \omega ||\boldsymbol{t} - \boldsymbol{t}'||^2 \Big)^{-\alpha}, \quad \alpha, \omega \geq 0.$$

Matérn:

$$k(\boldsymbol{t}, \boldsymbol{t}') = \frac{1}{\Gamma(\nu) 2^{\nu-1}} \Big(\sqrt{2\nu}\omega ||\boldsymbol{t} - \boldsymbol{t}'|| \Big)^{\nu} \mathcal{K}_{\nu} \Big(\sqrt{2\nu}\omega ||\boldsymbol{t} - \boldsymbol{t}'|| \Big), \quad \omega \geq 0,$$

where $\mathcal{K}_{\nu}$ is the modified Bessel function of order ν.

These kernels only depend on the Euclidean distance $d = ||\boldsymbol{t} - \boldsymbol{t}'||$.

More general norms

- to allow anisotropic covariance functions:

$$
\begin{aligned}
d^2 &= (\boldsymbol{t} - \boldsymbol{t}')^T \mathrm{diag}(\omega_1, \ldots, \omega_Q)(\boldsymbol{t} - \boldsymbol{t}') \\
&= \sum_{q=1}^{Q} \omega_q (t_q - t_q')^2, \qquad \omega_1, \ldots, \omega_Q \geq 0.
\end{aligned}
$$

- to allow non-separable covariance functions:

$$
d^2 = (\boldsymbol{t} - \boldsymbol{t}')^T \Sigma (\boldsymbol{t} - \boldsymbol{t}'), \qquad \text{where } \Sigma \text{ is positive semi-definite.}
$$

Considering nonstationarity

- When Q is small, $k(\cdot, \cdot)$ can be modelled nonparametrically (see e.g. Hall, Müller & Yao, 2008).
- When Q is large, nonparametric method suffers from the curse of dimensionality.

Considering nonstationarity

- When Q is small, $k(\cdot, \cdot)$ can be modelled nonparametrically (see e.g. Hall, Müller & Yao, 2008).
- When Q is large, nonparametric method suffers from the curse of dimensionality.
- We may use a parametric approach via a convolution (Higdon et al, 99):

$$f(\boldsymbol{t}) = \int_{\Re^2} k_{\boldsymbol{t}}(\boldsymbol{u})\psi(\boldsymbol{u})d\boldsymbol{u},$$

- Using a Gaussian kernel leads to (Paciorek and Schervish, 2006; Risser and Calder, 2017)

$$\mathrm{Cov}\left[f(\boldsymbol{t}), f(\boldsymbol{t}')\right] = \sigma^2 |\Sigma(\boldsymbol{t})|^{1/4} |\Sigma(\boldsymbol{t}')|^{1/4} \left| \frac{\Sigma(\boldsymbol{t}) + \Sigma(\boldsymbol{t}')}{2} \right|^{-1/2} g\left(\sqrt{Q_{\boldsymbol{t}\boldsymbol{t}'}}\right),$$

where g is a valid correlation function where

$$Q_{\boldsymbol{t}\boldsymbol{t}'} = (\boldsymbol{t} - \boldsymbol{t}')^T \left(\frac{\Sigma(\boldsymbol{t}) + \Sigma(\boldsymbol{t}')}{2}\right)^{-1} (\boldsymbol{t} - \boldsymbol{t}'),$$

Considering nonstationarity

- A special case is (composite GP, Ba and Joseph, 2012) that $\Sigma(\boldsymbol{t}) = \sigma(\boldsymbol{t})\boldsymbol{\Sigma}$, so that

$$\mathrm{Cov}\Big[f(\boldsymbol{t}), f(\boldsymbol{t}')\Big] = \sigma(\boldsymbol{t})\sigma(\boldsymbol{t}')|\Sigma|^{1/4}|\Sigma|^{1/4}\left|\frac{\Sigma+\Sigma}{2}\right|^{-1/2} g\left(\sqrt{Q_{\boldsymbol{t}\boldsymbol{t}'}}\right).$$

Considering nonstationarity

- A special case is (composite GP, Ba and Joseph, 2012) that $\Sigma(\boldsymbol{t}) = \sigma(\boldsymbol{t})\boldsymbol{\Sigma}$, so that

$$\mathrm{Cov}\left[f(\boldsymbol{t}), f(\boldsymbol{t}')\right] = \sigma(\boldsymbol{t})\sigma(\boldsymbol{t}')|\Sigma|^{1/4}|\Sigma|^{1/4}\left|\frac{\Sigma+\Sigma}{2}\right|^{-1/2} g\left(\sqrt{Q_{\boldsymbol{t}\boldsymbol{t}'}}\right).$$

- A general case: how to model $\Sigma(\boldsymbol{t})$ (Konzen, Shi and Wang, 2019)

Spherical parametrisation of varying matrix $\Sigma(\boldsymbol{\tau})$

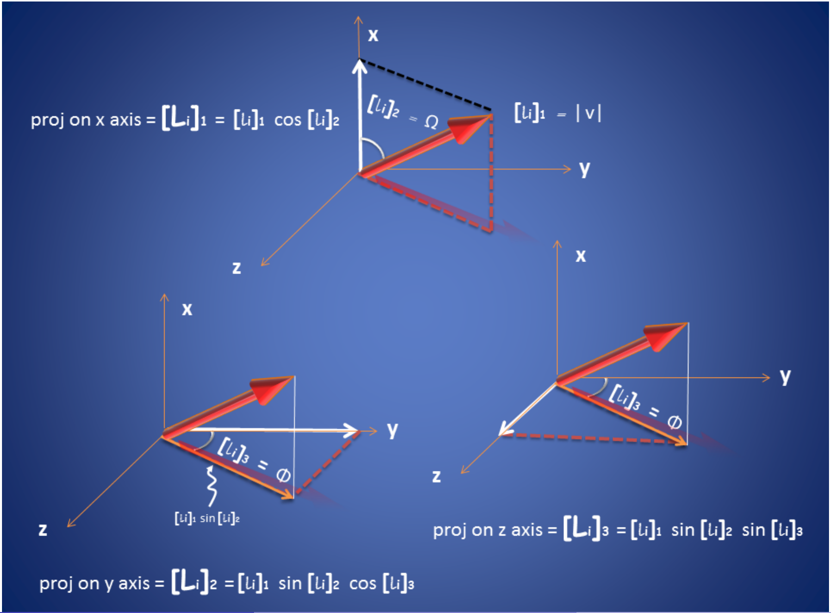
- $\boldsymbol{\tau}$ is a subset of $\boldsymbol{t}$
- We propose to use spherical parametrisation (Pinheiro and Bates, 1996) of $\Sigma(\boldsymbol{\tau})$, converting the problem to modelling of unconstrained parameters $\boldsymbol{\omega}(\boldsymbol{\tau})$.
- We will consider the Cholesky decomposition

$$\Sigma(\boldsymbol{\tau}) = L(\boldsymbol{\tau})^T L(\boldsymbol{\tau}),$$

where $L = L(\boldsymbol{\theta})$ is an $Q \times Q$ upper triangular matrix (including the main diagonal).

- Let L_i denote the ith column of L and ℓ_i denote the spherical coordinates of the first i elements of L_i.

Spherical parametrisation of varying matrix $\Sigma(\tau)$



Spherical parametrisation of varying matrix $\Sigma(\boldsymbol{\tau})$

- In general, we have

$$
\begin{aligned}
[L_i]_1 &= [\ell_i]_1 \cos([\ell_i]_2),\\
[L_i]_2 &= [\ell_i]_1 \sin([\ell_i]_2) \cos([\ell_i]_3),\\
&\ldots,\\
[L_i]_{i-1} &= [\ell_i]_1 \sin([\ell_i]_2) \cdots \cos([\ell_i]_i),\\
[L_i]_i &= [\ell_i]_1 \sin([\ell_i]_2) \cdots \sin([\ell_i]_i).
\end{aligned}
$$

- The spherical parameterisation is unique if

$$
\begin{aligned}
&[\ell_i]_1 > 0, \quad i = 1, \ldots, Q,\\
&[\ell_i]_j \in (0, \pi), \quad i = 2, \ldots, Q, \quad j = 2, \ldots, i.
\end{aligned}
$$

- Interpretation: we can show that $\Sigma_{ii} = [\ell_i]_1^2$ and that $\rho_{1i} = \cos([\ell_i]_2)$, $i = 2, \ldots, Q$, with $-1 < \rho_{1i} < 1$. This means that we can interpret the values of L in terms of the length-scale parameters and directions of dependence of Σ.

Nonstationary covaraince with varying matrix – Local empirical Bayesian estimation

- We can proceed with an unconstrained estimation by

$$\omega_i = \log([\ell_i]_1), \quad i = 1, \ldots, Q,$$

$$\omega_{Q+(i-2)(i-1)/2+j-1} = \log\left(\frac{[\ell_i]_j}{\pi - [\ell_i]_j}\right), \quad i = 2, \ldots, Q, \quad j = 2, \ldots, i.$$

- Model each $\omega_k(\boldsymbol{\tau})$ nonparametrically: e.g. by GPR or a set of basis functions

Nonstationary covaraince with varying matrix – Local empirical Bayesian estimation

- We can proceed with an unconstrained estimation by

$$\omega_i = \log([\ell_i]_1), \quad i = 1, \dots, Q,$$
$$\omega_{Q+(i-2)(i-1)/2+j-1} = \log\left(\frac{[\ell_i]_j}{\pi - [\ell_i]_j}\right), \quad i = 2, \dots, Q, \quad j = 2, \dots, i.$$

- Model each $\omega_k(\boldsymbol{\tau})$ nonparametrically: e.g. by GPR or a set of basis functions
- Then, we estimate the unconstrained hyperparameters $(\log \sigma_\varepsilon^2, \boldsymbol{\omega}(\boldsymbol{\tau}))$ via local marginal likelihood (or local empirical Bayesian), i.e. based on the marginal distribution of $X_n = \left(X(\boldsymbol{t}_1), \dots, X(\boldsymbol{t}_n)\right)'$.
- Flexible varying structure: e.g. time-varying or spatial-varying or both.
- Challenges: for non-Gaussian data

Prediction and decomposition of function-valued processes

- For Gaussian data, the posterior distribution $p(\boldsymbol{f}|\mathcal{D}, \sigma_\varepsilon^2)$ is a multivariate Gaussian distribution with

$$\hat{\boldsymbol{f}} = E(\boldsymbol{f}|\mathcal{D}, \sigma_\varepsilon^2) = K(K + \sigma_\varepsilon^2 I)^{-1}\boldsymbol{x}$$
$$\mathrm{Var}(\boldsymbol{f}|\mathcal{D}, \sigma_\varepsilon^2) = \sigma_\varepsilon^2 K(K + \sigma_\varepsilon^2 I)^{-1}.$$

Prediction and decomposition of function-valued processes

- For Gaussian data, the posterior distribution $p(\boldsymbol{f}|\mathcal{D},\sigma_\varepsilon^2)$ is a multivariate Gaussian distribution with

$$\hat{\boldsymbol{f}} = E(\boldsymbol{f}|\mathcal{D},\sigma_\varepsilon^2) = K(K+\sigma_\varepsilon^2 I)^{-1}\boldsymbol{x}$$
$$\mathrm{Var}(\boldsymbol{f}|\mathcal{D},\sigma_\varepsilon^2) = \sigma_\varepsilon^2 K(K+\sigma_\varepsilon^2 I)^{-1}.$$

- Decomposition (fPCA)

$$\begin{aligned} X(\boldsymbol{t}) &\approx \mu(\boldsymbol{t}) + \hat{f} \\ &= \mu(\boldsymbol{t}) + \sum_{j=1}^{\infty} \xi_j \phi_j(\boldsymbol{t}) \\ &\approx \mu(\boldsymbol{t}) + \sum_{j=1}^{J} \xi_j \phi_j(\boldsymbol{t}) \end{aligned}$$

GPR model – asymptotic theory

- Suppose that $k(\cdot,\cdot)$ continuous and has a finite trace, then $f(\boldsymbol{t})$ has a representation

$$f(\boldsymbol{t}) = \sum_{j=1}^{\infty} \xi_j \phi_j(\boldsymbol{t}) = \sum_{j=1}^{J} \xi_j \phi_j(\boldsymbol{t}) + b^{1/2} z(\boldsymbol{t})$$

where $\lambda_1 \geq \lambda_2 \ldots$, and ϕ_j is the eigen-function of $k(\cdot,\cdot)$ and $\xi_j \sim N(0, \lambda_j)$.

- We therefore have RKHS

$$\mathcal{H}_K = \mathcal{H}_0 \oplus \mathcal{H}_1,$$

where $\mathcal{H}_0$ is the span of $\phi_1, \cdots, \phi_S$ (null space) and $\mathcal{H}_1$ is the RKHS for K_1.

- Let $\mathcal{P}_1$ be the orthogonal projection operator in $\mathcal{H}_K$ onto $\mathcal{H}_1$, and $f_{n,\lambda}$ be the nimimiser in $\mathcal{H}_K$ of the regularised risk functional:

$$\frac{1}{n} \sum_{i=1}^{n} (x_i - f(\boldsymbol{t}_i))^2 + \lambda ||\mathcal{P}_1 f||_K,$$

GPR model – asymptotic theory

Theorem

Let $\hat{f}_{GP}(\boldsymbol{t}) = E(f(\boldsymbol{t})|x_1, \ldots, x_n)$*, then*

$$\lim_{D \to \infty} \hat{f}_{GP}(\boldsymbol{t}) = f_{n,\lambda}(\boldsymbol{t}),$$

where $\lambda = \frac{\sigma^2}{nb}$ *and* $\boldsymbol{D} = diag(\lambda_1/b, \ldots, \lambda_S/b)$*.* $\lim_{\boldsymbol{D} \to \infty}$ *means that each element tends to infinity.*

GPR model: posterior consistency

Theorem

(Choi, 2005) Let P_0 denote the joint conditional distribution of $\{x_n\}_{n=1}^{\infty}$ given the covariate assuming that f_0 is the true response function. Suppose that the values of the covariate in $[0,1]$ are fixed, i.e., known ahead of time. Then for every $\epsilon > 0$,

$$\Pi\left\{f \in W_{\epsilon,n}^{C} | \mathcal{D}\right\} \to 0 \ a.s.\ [P_0].$$

The neighbourhood is defined as

$$W_{\epsilon,n} = \left\{(f,\sigma) \ : \ \int |f(\boldsymbol{t}) - f_0(\boldsymbol{t})| dQ_n(x) < \epsilon, \ \left|\frac{\sigma}{\sigma_0} - 1\right| < \epsilon\right\}.$$

GPR model: posterior consistency

Theorem

(Choi, 2005) Let P_0 denote the joint conditional distribution of $\{x_n\}_{n=1}^{\infty}$ given the covariate assuming that f_0 is the true response function. Suppose that the values of the covariate in $[0, 1]$ are fixed, i.e., known ahead of time. Then for every $\epsilon > 0$,

$$\Pi\left\{f \in W_{\epsilon,n}^{C} | \mathcal{D}\right\} \rightarrow 0 \ a.s. \ [P_0].$$

The neighbourhood is defined as

$$W_{\epsilon,n} = \left\{(f, \sigma) \ : \ \int |f(\boldsymbol{t}) - f_0(\boldsymbol{t})| dQ_n(x) < \epsilon, \ \left|\frac{\sigma}{\sigma_0} - 1\right| < \epsilon\right\}.$$

Remarks: a good choice of hyper-parameters can improve the efficiency, but has no influence to the consistency

GPR model: information consistency

- K-L distance: $D[p, q] = \int (\log p - \log q) dP$.

Theorem

Upper bound of $D[P_0(x_1, \ldots, x_n | f_0), P_{GP}(x_1, \ldots, x_n)]$,

$$D[P_0(x_1, \ldots, x_n | f_0), P_{GP}(x_1, \ldots, x_n)] \le \frac{1}{2} \|f_0\|_{\boldsymbol{K}}^2 + \frac{1}{2} \log |\boldsymbol{I}_n + c\boldsymbol{K}|,$$

- $\|f\|_{\boldsymbol{K}}$ *is the RKHS norm of* f, *and* c *is a certain constant.*
- $P_{GP}(x_1, \ldots, x_n)$ *– a Bayesian predictive distribution of* $x_1, \ldots, x_n$ *using GP prior based on* n *observations.*

- Thus, the expected KL divergence divided by the sample size converges to zero as the sample size increases (Seeger, et al. 2008).

Decomposition of function-valued processes – Asymptotic theory

Theorem

For $N \geq 1$ for which $\lambda_N > 0$, functions $\{\phi_i, i = 1, ..., N\}$ provide the best finite dimensional approximations to $Z^c(\boldsymbol{u})$ with respect to minimizing criterion

$$\operatorname{argmin}_{g_1,\ldots,g_N \in L^2(\mathcal{U})} E\left\{ \int_{\mathcal{U}} ||Z^c(\boldsymbol{u}) - \sum_{i=1}^{N} g_i(\boldsymbol{u})\xi_i^*||^2 d\boldsymbol{u} \right\},$$

where $g_1, ..., g_N \in L^2(\mathcal{U})$ are orthogonal, and $\xi_i^* =< Z^c(\cdot), g_i(\cdot) >= \int Z^c(\boldsymbol{u})g_i(\boldsymbol{u})du$. The minimizing value is $\sum_{i=N+1}^{\infty} \lambda_i$.

Decomposition of function-valued processes – Asymptotic theory

Theorem

For $N \geq 1$ for which $\lambda_N > 0$, functions $\{\phi_i, i = 1, ..., N\}$ provide the best finite dimensional approximations to $Z^c(\boldsymbol{u})$ with respect to minimizing criterion

$$\operatorname{argmin}_{g_1,\ldots,g_N \in L^2(\mathcal{U})} E\left\{\int_{\mathcal{U}} ||Z^c(\boldsymbol{u}) - \sum_{i=1}^{N} g_i(\boldsymbol{u})\xi_i^*||^2 d\boldsymbol{u}\right\},$$

where $g_1, ..., g_N \in L^2(\mathcal{U})$ are orthogonal, and $\xi_i^* =< Z^c(\cdot), g_i(\cdot) >= \int Z^c(\boldsymbol{u})g_i(\boldsymbol{u})du$. The minimizing value is $\sum_{i=N+1}^{\infty} \lambda_i$.

Theorem

Suppose conditions C1 - C3 in Appendix hold, and $\hat{\mu}(\boldsymbol{t})$ satisfies $\sup_{\boldsymbol{t}} |\hat{\mu}(\boldsymbol{t}) - \mu(\boldsymbol{t})| = O_p[\{\log(n)/n\}^{1/2}]$, *we have, for $1 \leq i \leq N$,*

$$||k_{\hat{\theta}}(\cdot,\cdot) - k_\theta(\cdot,\cdot)|| = O_p(\{\log(n)/n\}^{1/2}),$$
$$||\hat{\lambda}_i - \lambda_i|| = O_p(\{\log(n)/n\}^{1/2}),$$
$$||\hat{\phi}_i(\cdot) - \phi_i(\cdot)|| = O_p(\{\log(n)/n\}^{1/2}),$$
$$||\hat{\xi}_i - \xi_i|| = O_p(\{\log(n)/n\}^{1/2}).$$

An example using a general covariance structure

In this simulation study, we assume that the random process $f(t_1, t_2)$ has zero mean and covariance function given by

$$\mathrm{Cov}\Big[f(t_1, t_2), f(t_1', t_2')\Big] = \sum_{j=1}^{20} \alpha_j \phi_j(t_1 + t_2)\phi_j(t_1' + t_2'),$$

where $\phi_j(\cdot)$ are Chebyshev polynomials, $\alpha_j = j^{-3/2}$ and $\boldsymbol{t} \in [-1, 1]^2$.

We have generated 100 curves from $X(\boldsymbol{t}) = f(\boldsymbol{t}) + \varepsilon$, $\sigma_\varepsilon^2 = 0.1^2$, observed at $n_1 \times n_2 = 20 \times 20 = 400$ equally spaced points.

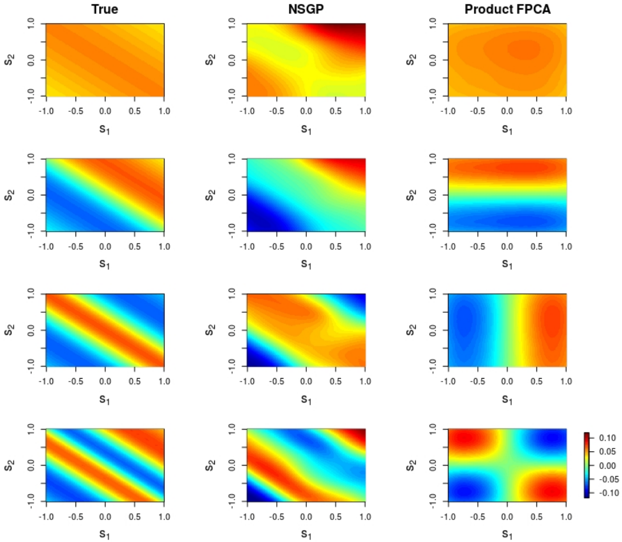


Figure 2: First four leading eigensurfaces $\phi(t_1, t_2)$ of the true model (left column) and the corresponding estimated eigensurfaces $\hat{\phi}(t_1, t_2)$ from the nonstationary GP model (centre) and Product FPCA model (right).

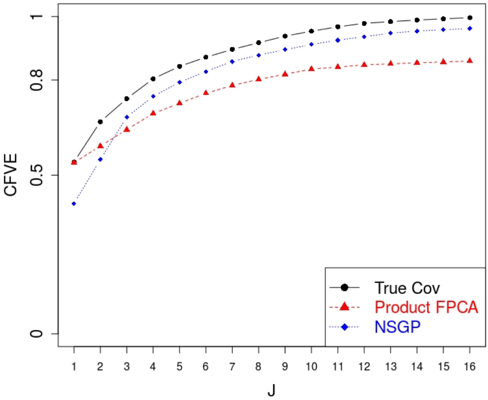
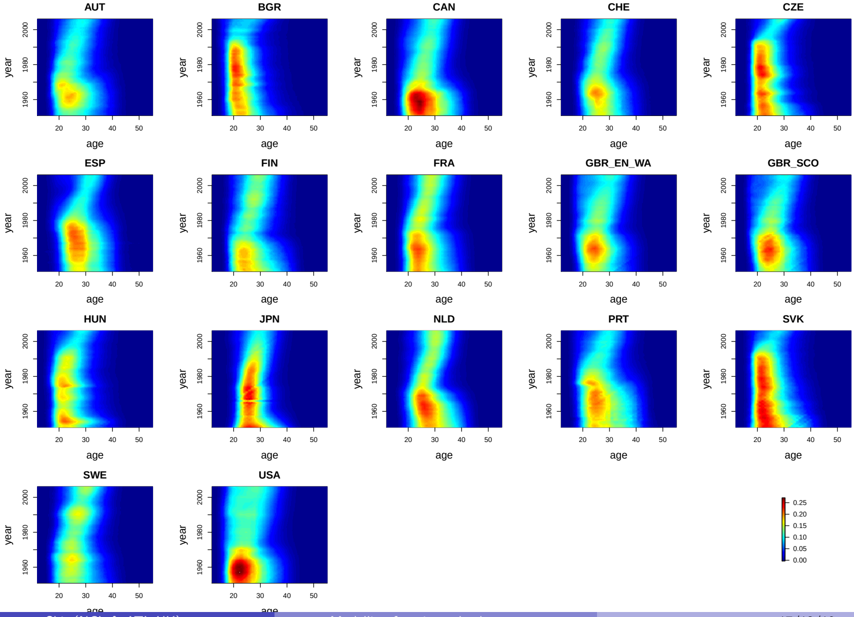


Figure 3: Comparison of cumulative FVEs obtained by the true, and Product FPCA, and nonstationary GP (NSGP) models.

Application 1: Non-stationary Gaussian Processes applied to ASFR data



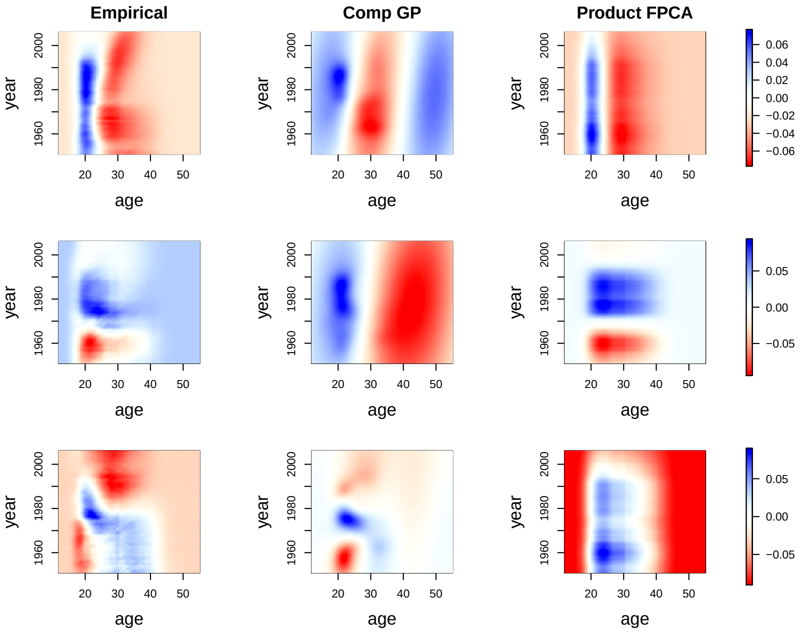


Figure 5: First three eigensurfaces $\hat{\phi}_j(s,t)$, $j = 1, 2, 3$, of the Empirical, Composite GP, and Product FPCA covariance functions estimated for ASFR of 17 countries.

Conclusions

- By avoiding the covariance separability assumption, we can provide additional insights into multi-dimensional functional data;
- Extensions to cases where $Q > 2$ are straightforward;
- We just need one realisation of the random process X to estimate its covariance structure;
- Convolved GPs can be used to measure the cross-covariance structure between functions.

Conclusions

- By avoiding the covariance separability assumption, we can provide additional insights into multi-dimensional functional data;
- Extensions to cases where $Q > 2$ are straightforward;
- We just need one realisation of the random process X to estimate its covariance structure;
- Convolved GPs can be used to measure the cross-covariance structure between functions.
- Interesting topics for future research
  - Extension to multi-variate function-valued processes, i.e. $X(\boldsymbol{t}) \in \mathbb{R}^m, \ \boldsymbol{t} \in \mathbb{R}^Q$

Conclusions

- By avoiding the covariance separability assumption, we can provide additional insights into multi-dimensional functional data;
- Extensions to cases where $Q > 2$ are straightforward;
- We just need one realisation of the random process X to estimate its covariance structure;
- Convolved GPs can be used to measure the cross-covariance structure between functions.
- Interesting topics for future research
  - Extension to multi-variate function-valued processes, i.e. $X(\boldsymbol{t}) \in \mathbb{R}^m, \ \boldsymbol{t} \in \mathbb{R}^Q$
  - The use of other process priors: e.g. heavy-tailed processes (Shah et al., 2014; Wang et al., 2017; Cao et al., 2018): need efficient algorithm
  - Extension to Non-Gaussian data is challenging.

Conclusions

- By avoiding the covariance separability assumption, we can provide additional insights into multi-dimensional functional data;
- Extensions to cases where $Q > 2$ are straightforward;
- We just need one realisation of the random process X to estimate its covariance structure;
- Convolved GPs can be used to measure the cross-covariance structure between functions.
- Interesting topics for future research
  - Extension to multi-variate function-valued processes, i.e. $X(\boldsymbol{t}) \in \mathbb{R}^m$, $\boldsymbol{t} \in \mathbb{R}^Q$
  - The use of other process priors: e.g. heavy-tailed processes (Shah et al., 2014; Wang et al., 2017; Cao et al., 2018): need efficient algorithm
  - Extension to Non-Gaussian data is challenging.

Thanks for listening!

References

Chen, K., Delicado, P., and Müller, H.-G. (2017). Modelling function-valued stochastic processes, with applications to fertility dynamics. *J. R. Statist. Soc. B*, 79(1):177–196.

Konzen, E., Shi, J. Q. and Wang, Z. (2019). Modelling function-valued processes with nonseparable covariance structure. Technical report, Newcastle University.

Paciorek, C. J. and Schervish, M. J. (2006). Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506.

Risser, M. and Calder, C. (2017). Local likelihood estimation for covariance functions with spatially-varying parameters: The convoSPAT package for R. *Journal of Statistical Software*, Articles, 81(14):1–32.

Shi, J. Q. and Choi, T. (2011). *Gaussian process regression analysis for functional data*. CRC Press.

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