The Elements of Gaussian process

Jianbin Tan

School of Mathematics
Sun Yat-sen University

Table of Contents

Introduction

Statistical inference

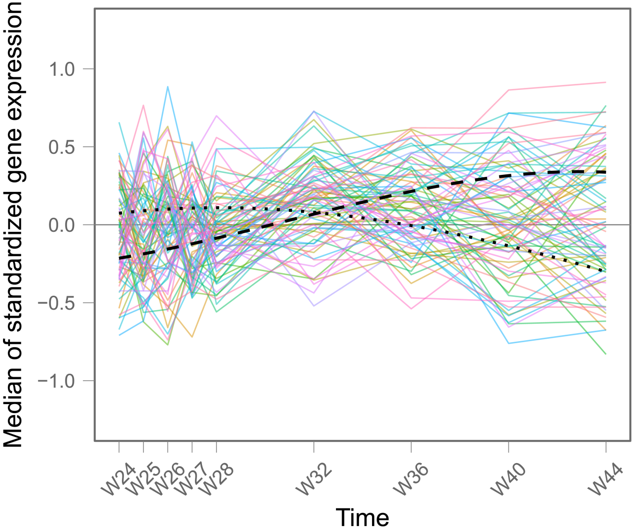
- Stationary covariance functions
  - Random covariance function
  - Spectral density induced covariance function
  - Stochastic differential equation induced covariance function
- Non-stationary Covariance Functions
  - Differential operator induced covariance function
  - Neural network induced covariance function
- Non-parametric covariance function

Further examples

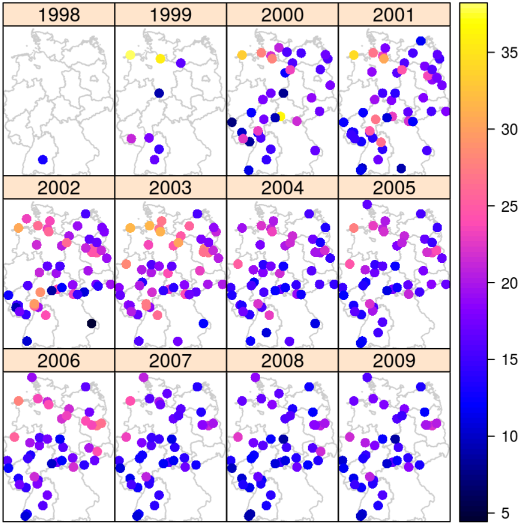
Gaussian processes are versatile tools for statistical modeling of random functions.

- Observable processes, e.g. physical / biological processes, spatial / temporal / spatiotemporal processes.
- Unobservable processes, e.g. latent processes, point processes.
- In Bayesian statistics, all the functional parameters are assumed to be random.

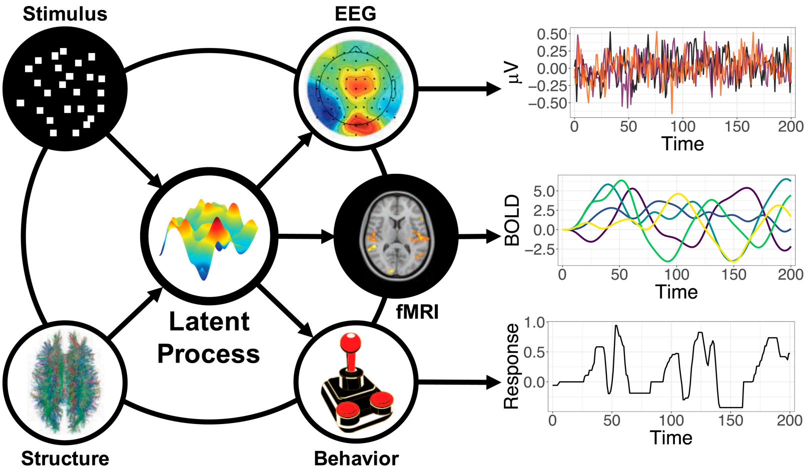
Biological processes



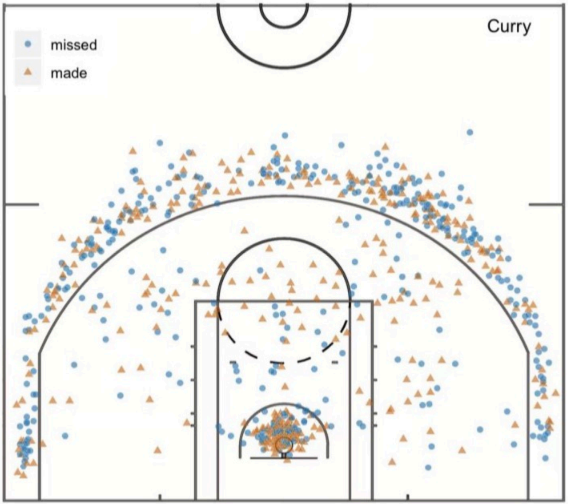
Spatiotemporal processes



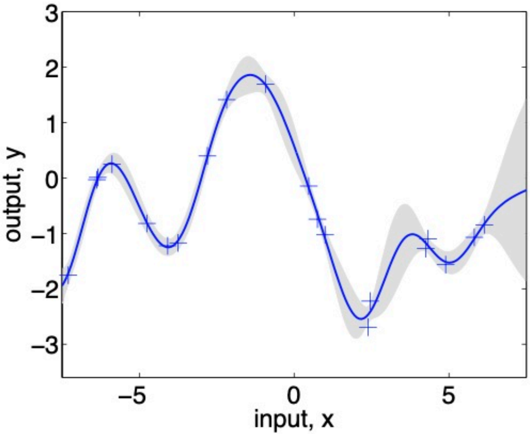
Latent processes



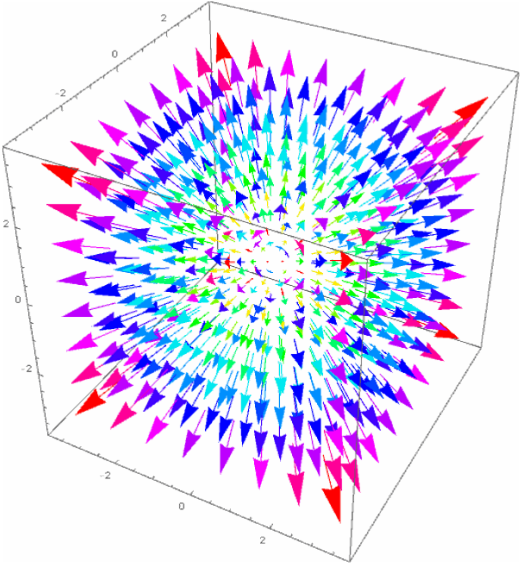
Point processes



Functional approximation



Vector field



Definition

X_t is a random variable indexed by $t \in \mathcal{T}$, usually, $\mathcal{T}$ is a continuum (but is not necessary). We call that $(X_t)_{t\in\mathcal{T}}$ is a Gaussian processes if $(X_t)_{t\in\mathcal{T}_d}$ follows the multivariate Gaussian distribution for all finite grid $\mathcal{T}_d \subset \mathcal{T}$.

- The existence of Gaussian processes is guaranteed by Kolmogorov extension theorem.
- By the properties of Gaussian distribution, $(X_t)_{t\in\mathcal{T}}$ is fully determined by

$$\begin{aligned} \mu(t) &:= \mathbb{E}X_t, \\ K(t,s) &:= \mathrm{Cov}(X_t, X_s), \end{aligned}$$

  which are called the mean function and covariance function of the Gaussian process.
- We call denote X_t as $X(t)$ to emphasize its functional nature, for this case, $X(\cdot)$ can be viewed as a random element under certain regularization conditions.

Mean-square continuous processes

We call that $(X_t)_{t \in \mathcal{T}}$ is a mean-square continuous processes if

$$\lim_{t_n \to t} \mathbb{E}(X_{t_n} - X_t)^2 = 0,$$

$\forall \ t_n \to t$. When $\mathbb{E}X_t^2 < \infty$ for all t, $(X_t)_{t \in \mathcal{T}}$ is a mean-square continuous processes if and only if $\mu(\cdot)$ and $K(\cdot)$ are continuous.

- We prefer to use continuous covariance kernel $K(\cdot)$ to obtain a well behaved realization of Gaussian processes.
- When there exists α, β, and C s.t.

$$\mathbb{E}|X_s - X_t|^\alpha \leq C|s - t|^{1+\beta},$$

$(X_t)_{t \in \mathcal{T}}$ has a continuous modification by Kolmogorov's criterion.

Derivative of Gaussian process

We define that

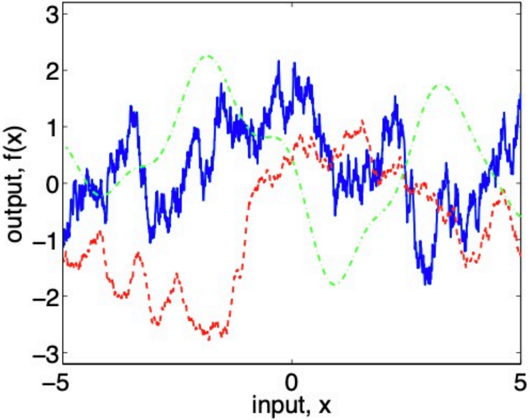
$$X'(t) := \lim_{h\to 0} \frac{X(t+h)-X(t)}{h},$$

when the R.H.S of the above limit exists in the mean square sense. Noting that

$$\begin{aligned}\mathbb{E}\left(\frac{X(t+h)-X(t)}{h}\right)^2 &= \frac{K(t+h,t+h)-K(t,t+h)-(K(t,t+h)-K(t,t))}{h^2}\\ &\to \frac{dK(t,t)}{dt^2}\end{aligned}$$

when $\frac{dK(t,t)}{dt^2}$ exists. The existence of the second-order derivative of $K(\cdot)$ plays an essential role in determining the existence of the derivative of $X(\cdot)$.

Gaussian processes with different covariance kernels



Mean function

To infer a Gaussian process $(X_t)_{t \in \mathcal{T}}$, our goal is to estimate $\mu(t)$ and $K(t, s)$. We first focus a common observational scheme. We assume that we observe

$$X(t_1) + \varepsilon_1,\ X(t_2) + \varepsilon_2, \cdots, X(t_n) + \varepsilon_n,$$

where ε_i's are mean-zero noises. For this case, we may estimate $\mu(\cdot)$ by using a para -metric or non-parametric representation of $\mu(\cdot)$ to fit the observed data. But mostly, the mean function is not our primary goal under this observational design. In order to infer $X(\cdot)$ itself, we usually assume that $\mu(t) = 0$.

Mean-zero Gaussian process

When we use Gaussian process to represent functional object. A fundamental problem is that

How about the model complexity of a Gaussian process?

As we have assumed that $\mu(t) = 0$, the answer is hidden within the covariance function $K()$.

Mercer's theorem

Assuming $K(\cdot)$ is continuous on a compact set, then

$$K(t, s) = \sum_{k=1}^{\infty} \eta_k \phi_k(t) \phi_k(s),$$

where η_k is the eigenvalue and $\phi_k(\cdot)$ is the normalized eigenfunction. This lead to the Karhunen-Loeve expansion of $X(\cdot)$ as

$$X(t) = \sum_{k=1}^{\infty} a_k \phi_k(t),$$

where a_k is a mean-zero Gaussian variable with variance η_k.

Eigenfunctions

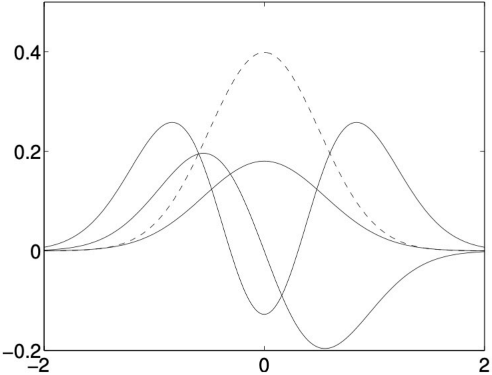


Figure: The first 3 eigenfunctions of the squared exponential kernel $K(t, s) = \exp(-(t-s)^2)$ w.r.t. a Gaussian density.

Parametric estimates of covariance functions

We now assume that $\mu(t) = 0$ and $Y_i = X(t_i) + \varepsilon_i$, $i = 1, \cdots, n$, and we parame -terize K as K_α. Additional, we assume that ε_i's is i.i.d. mean-zero Gaussian variables with variance σ^2. This model setting is commonly used when n is small and t_i's are irregular observed. Note that

$$\boldsymbol{Y} \sim \text{Gau}(\boldsymbol{0}, \boldsymbol{K}_\alpha + \sigma^2 \boldsymbol{I}),$$

where $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ and $\boldsymbol{K}_\alpha = (K_\alpha(t_i, t_j))_{i,j=1,\cdots,n}$. The MLE estimator can be directly established by maximizing σ^2 and α in the observed likelihood.

Conditional expectation

Given estimates of σ^2 and α, we approximate $X(t)$ by conditional expectation. Notice that

$$X(t) \mid \boldsymbol{Y} \sim \mathrm{Gau}(\mathcal{K}_\alpha(t, \mathbb{T}_o)\ (\boldsymbol{K}_\alpha + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{Y},\ K(t,t) - \mathcal{K}_\alpha(t, \mathbb{T}_o)\ (\boldsymbol{K}_\alpha + \sigma^2 \boldsymbol{I})^{-1} \mathcal{K}_\alpha(\mathbb{T}_o, t)),$$

where $\mathbb{T}_o = \{t_1, \cdots, t_n\}$.

- The inference using the conditional distribution is optimal under the decision theory framework.
- $\mathcal{K}_\alpha(t, \mathbb{T}_o)\ (\boldsymbol{K}_\alpha + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{Y}$ is also called the MAP (Maximum a posteriori) estimator of $X(t)$.
- $\mathcal{K}_\alpha(t, \mathbb{T}_o)\ (\boldsymbol{K}_\alpha + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{Y} = h^T(t) \boldsymbol{Y}$ is a linear smoother.

Conditional expectation

Given estimates of σ^2 and α, we approximate $X(t)$ by conditional expectation. Notice that

$$X(t) \mid \boldsymbol{Y} \sim \mathrm{Gau}(\mathcal{K}_\alpha(t, \mathbb{T}_o)\ (\boldsymbol{K}_\alpha + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{Y},\ K(t,t) - \mathcal{K}_\alpha(t, \mathbb{T}_o)\ (\boldsymbol{K}_\alpha + \sigma^2 \boldsymbol{I})^{-1} \mathcal{K}_\alpha(\mathbb{T}_o, t)),$$

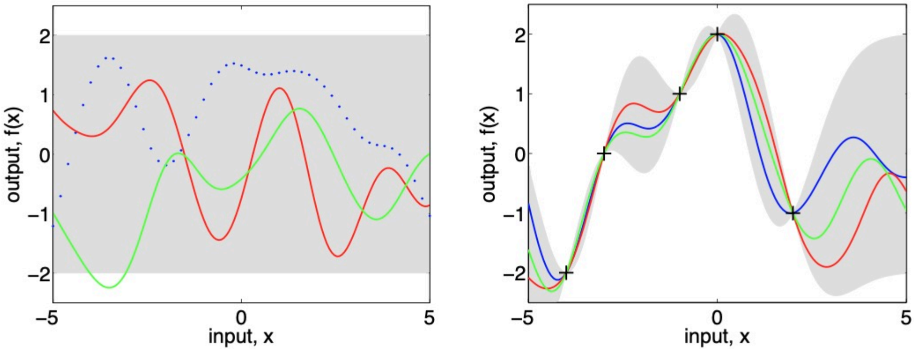
where $\mathbb{T}_o = \{t_1, \cdots, t_n\}$.

- $\mathcal{K}_\alpha(t, \mathbb{T}_o)\ (\boldsymbol{K}_\alpha + \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{Y} = \sum_{i=1}^n \hat{\beta}_i K_\alpha(t, t_i)$ is the kernel ridge regression estimator of $X(t)$, i.e.

$$\min_{X(\cdot) \in \mathcal{H}(K_\alpha)} \frac{\|\boldsymbol{Y} - X(\mathbb{T}_o)\|^2}{2\sigma^2} + \|X(\cdot)\|^2_{\mathcal{H}(K_\alpha)},$$

where $\mathcal{H}(K_\alpha)$ is the reproducing kernel Hilbert space (RKHS) with kernel $K_\alpha(\cdot)$ and norm $\|\cdot\|_{\mathcal{H}(K_\alpha)}$.

Conditional distribution



(a), prior (b), posterior

Main problems

- Although Gaussian are convenient from both an analytical and a practical point of view, the computational issues have always been a bottleneck. This is due to the general cost of $O(n^3)$ to factorize dense $n \times n$ (covariance) matrices.
- How to chose the parametric kernel $K_\alpha(\cdot)$.

Table of Contents

Introduction

Statistical inference

Stationary covariance functions

Random covariance function

Spectral density induced covariance function

Stochastic differential equation induced covariance function

Non-stationary Covariance Functions

Differential operator induced covariance function

Neural network induced covariance function

Non-parametric covariance function

Further examples

Stationary covariance functions

When $K(t,s) = K(t-s)$, we called that K is weak stationary. We have the next theorem.

(Bochner's theorem) A complex-valued function K on $\mathbb{R}^p$ is the covariance function of a weakly stationary mean square continuous complex valued random process on $\mathbb{R}^D$ if and only if it can be represented as

$$K(\boldsymbol{\tau}) = \int_{\mathbb{R}^D} e^{\mathrm{i}\boldsymbol{\omega}^T \boldsymbol{\tau}} d\mu(\boldsymbol{\omega})$$

where μ is a positive finite measure. If μ has a density S, then S is known as the spectral density.

Random feature

If the kernel $K(\boldsymbol{\tau})$ is properly scaled, Bochner's theorem guarantees that its Fourier transform $S(\boldsymbol{\omega})$ is a proper probability distribution. Defining $\zeta_{\boldsymbol{\omega}}(\mathbf{t}) = e^{j\boldsymbol{\omega}'\mathbf{t}}$, we have

$$K(\mathbf{t}-\mathbf{s}) = \int_{\mathbb{R}^d} S(\boldsymbol{\omega}) e^{j\boldsymbol{\omega}^T(\mathbf{t}-\mathbf{s})} d\boldsymbol{\omega} = \mathbb{E}\left[\zeta_{\boldsymbol{\omega}}(\mathbf{t})\zeta_{\boldsymbol{\omega}}(\mathbf{s})^*\right].$$

Therefore, we may specify a probability measure S and use Monte Carlo sampling to approximate the covariance K.

- The computation is fast when n is relatively large (separability).
- Save memory.

Whittle likelihood

Denote $Y_i = Y(t_i)$. Noting that

$$G_{\alpha,\sigma^2}(t_i, t_j) := \mathrm{Cov}(Y(t_i), Y(t_j)) = K_\alpha(t_i, t_j) + \sigma^2 \mathbb{I}(t_i = t_j)$$

is also stationary when K_α dose. Let S_{α,σ^2} be the spectral density of G_{α,σ^2}. Define $y(\theta) = \frac{1}{\sqrt{2\pi n}} \sum_{i=1}^{n} Y(t_i) e^{\mathrm{i} j\theta}$. It can be shown that

- $y(\theta)$ asymptotically follows a mean-zero Gaussian distribution with covariance $S_{\alpha,\sigma^2}(\theta)$.
- $y(\theta_1)$ and $y(\theta_2)$ are asymptotically independent.

Whittle likelihood

Then a pseudo likelihood of $\boldsymbol{Y}$ can be constructed as

$$\sum_{\theta \in [-\pi, \pi]} \log(|S_{\alpha,\sigma^2}(\theta)|) - |y(\theta)|^2 / S_{\alpha,\sigma^2}(\theta).$$

- Don't need to inverse a matrix.
- Require that t_i's are regularly spaced and densely observed.

Dynamic representation of Gaussian processes

Despite its generality, computation of Gaussian processes methodology does not conveniently:

- Large n case.
- Irregularly and sparsely observed data.

For this class of Gaussian processes, instead of considering the kernel formalism, it is appealing to work with its mathematical dual, where the models are written out in terms of dynamical systems.

Continuous state-space model

In the following, the interest will be in classes of covariance functions which can be represented in terms of a dynamical model (the Gaussian prior) and measurement model of the form:

$$\frac{\mathrm{d}\mathbf{X}(t)}{\mathrm{d}t} = \mathbf{F}\mathbf{X}(t) + \mathbf{L}\mathbf{w}(t),$$
$$\mathbf{Y}_k = \mathbf{H}\mathbf{X}\left(t_k\right) + \varepsilon_k,$$

where

- $\mathbf{X}(t) = (X_1(t), X_2(t), \cdots, X_m(t))$ contains the m stochastic processes.
- $\mathbf{w}(t) \in \mathbb{R}^s$ is a multi-dimensional white noise process with spectral density matrix $\mathbf{Q}_\mathrm{c} \in \mathbb{R}^{s \times s}$.
- The model is defined by the feedback matrix $\mathbf{F} \in \mathbb{R}^{m \times m}$, the noise effect matrix $\mathbf{L} \in \mathbb{R}^{m \times s}$, and the initial state is assumed to be a Gaussian vector with covariance $\mathbf{P}_0$.
- The measurements are assumed to be corrupted by i.i.d. Gaussian noise, $\varepsilon_k \sim \text{Gau}\left(0, \sigma^2\right)$.

Matérn covariance

The Matérn covariance can be constructed by the stochastic differential equation model above, which is

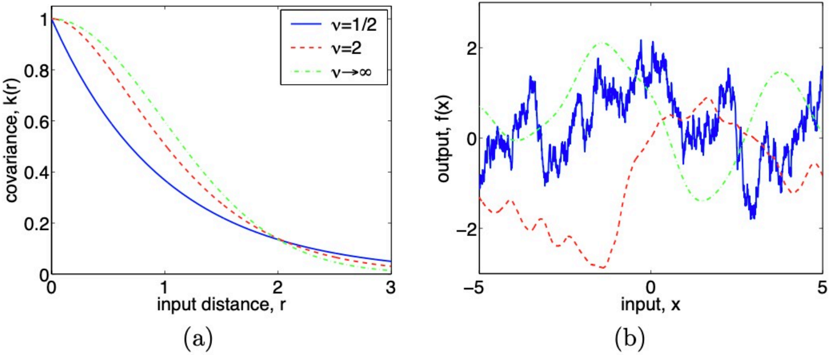
$$C(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{d}{\rho} \right)^{\nu} K_{\nu} \left(\sqrt{2\nu} \frac{d}{\rho} \right),$$

where Γ is the gamma function, K_ν is the modified Bessel function, and ρ and ν are positive parameters of the covariance. Its spectral density is

$$S(\omega) = \sigma^2 \frac{2^n \pi^{n/2} \Gamma\left(\nu + \frac{n}{2}\right) (2\nu)^{\nu}}{\Gamma(\nu) \rho^{2\nu}} \left(\frac{2\nu}{\rho^2} + 4\pi^2 \omega^2 \right).$$

A Gaussian process with Matérn covariance is $\lceil \nu \rceil - 1$ times differentiable in the mean-square sense.

Gaussian processes with different covariance kernels



Gaussian processes with different covariance kernels

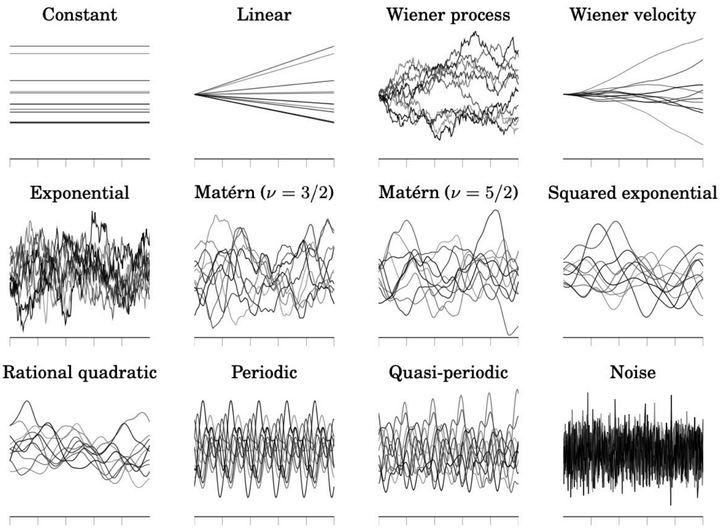


Table of Contents

Introduction

Statistical inference

Stationary covariance functions

Random covariance function

Spectral density induced covariance function

Stochastic differential equation induced covariance function

Non-stationary Covariance Functions

Differential operator induced covariance function

Neural network induced covariance function

Non-parametric covariance function

Further examples

Homogeneous differential equation

Given a differential operator $\mathcal{D}$, we may want to assume that

$$\int |\mathcal{D}X(t)|^2 \, \mathrm{d}t$$

is small stochastically, i.e., $\|X\|_{\mathcal{H}(K_{\mathcal{D}})}$ is small stochastically, where

$$K_{\mathcal{D}}(t_1, t_2) = \int G(t_1, s) G(t_2, s) \, \mathrm{d}s$$

with $G(\cdot)$ being a Green's function of $\mathcal{D}$.

Homogeneous differential constraint

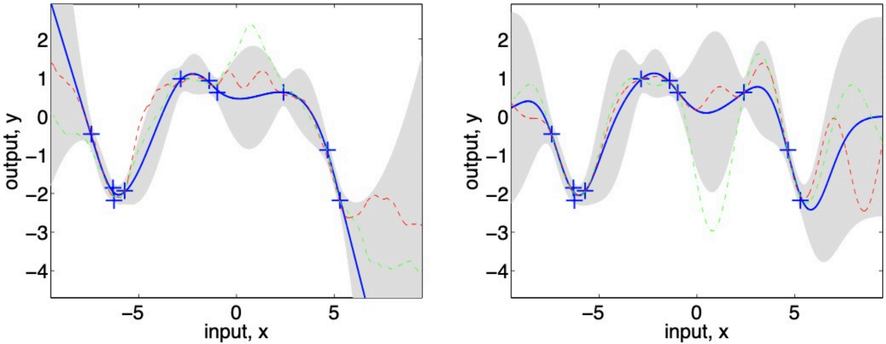
By the connection between RKHS and Gaussian processes, we may use $K_{\mathcal{D}}$ as the covariance function of a mean-zero Gaussian processes $X_1(\cdot)$, and assume that

$$X(t) = \mu(t) + X_1(t)$$

where $\mu(t)$ is the mean function of $X(\cdot)$ satisfying that $\mathcal{D}\mu(t) = 0$.

- $\mu(t)$ is usually contained in a finite-dimensional function space, and is needed to be estimated.
- We usually assume that $\mu(t)$ is also a mean-zero Gaussian processes under the Bayesian framework.

Inference



(a), spline covariance (b), squared exponential cov.

Neural network representation

We briefly review the correspondence between single-hidden layer neural networks and Gaussian process. Here, we usually focus on the $X()$ that the domain and image of which is high-dimensional, and re-define the random function as $\boldsymbol{z}^1(\boldsymbol{x})$. Moreover, the i th component of the network output, z_i^1, is parameterized as

$$z_i^1(\boldsymbol{x}) = b_i^1 + \sum_{j=1}^{N_1} W_{ij}^1 x_j^1(x), \quad x_j^1(\boldsymbol{x}) = \phi\left(b_j^0 + \sum_{k=1}^{d_{in}} W_{jk}^0 x_k\right)$$

We take the weight and bias parameters to be i.i.d. random variables, so that the post-activations $x_j^1, x_{j'}^1$ are independent for $j \neq j'$ conditioning on x.

Neural network representation

- Since $z_i^1(\boldsymbol{x})$ is a sum of i.i.d terms, it follows from the generalized Central Limit Theorem that in the limit of infinite width $N_1 \to \infty, z_i^1(\boldsymbol{x})$ will be Gaussian distributed.
- Likewise, from the multidimensional Central Limit Theorem, any finite collection of $\left\{z_i^1\left(\boldsymbol{x}^1\right), \ldots, z_i^1\left(\boldsymbol{x}^h\right)\right\}$ will have a joint multivariate Gaussian distribution, which is exactly the definition of a Gaussian process.
- Therefore we conclude that $z_i^1(\boldsymbol{x})$ follows a Gaussian process with mean μ^1 and covariance K^1, which are themselves independent of i conditioning on x.

Neural network representation

We assume that the parameters have zero mean, then $\mu^1(\boldsymbol{x}) = \mathbb{E}\left[z_i^1(\boldsymbol{x})\right] = 0$ and,

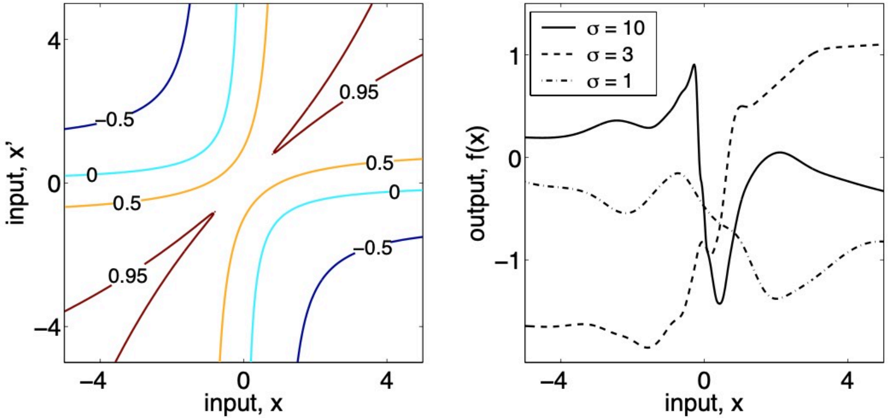
$$K^1\left(\boldsymbol{x}, \boldsymbol{x}'\right) \equiv \mathbb{E}\left[z_i^1(\boldsymbol{x}) z_i^1\left(\boldsymbol{x}'\right)\right] = \sigma_b^2 + \sigma_w^2 \mathbb{E}\left[x_i^1(\boldsymbol{x}) x_i^1\left(\boldsymbol{x}'\right)\right] := \sigma_b^2 + \sigma_w^2 C\left(\boldsymbol{x}, \boldsymbol{x}'\right),$$

We just need to specify the kernel C given prior distribution of hidden parameters and ϕ. One form of C can be constructed as

$$C\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \frac{2}{\pi} \sin^{-1}\left(\frac{2\tilde{\mathbf{x}}^T \Sigma \tilde{\mathbf{x}}'}{\sqrt{\left(1 + 2\tilde{\mathbf{x}}^\top \Sigma \tilde{\mathbf{x}}\right)\left(1 + 2\tilde{\mathbf{x}}'^\top \Sigma \tilde{\mathbf{x}}'\right)}}\right),$$

where $\tilde{\boldsymbol{x}} = (1, \boldsymbol{x})^T$.

Neural network kernel



(a), covariance (b), sample functions

Multiple layers

Suppose that z_j^{l-1} is a Gaussian process, identical and independent for every j. After $l-1$ steps, the network computes

$$z_i^l(\boldsymbol{x}) = b_i^l + \sum_{j=1}^{N_l} W_{ij}^l x_j^l(\boldsymbol{x}), \quad x_j^l(\boldsymbol{x}) = \phi\left(z_j^{l-1}(\boldsymbol{x})\right).$$

As before, $z_i^l(\boldsymbol{x})$ is a sum of i.i.d. random terms so that, as $N_l \to \infty$, any finite collection $\left\{z_i^l\left(\boldsymbol{x}^1\right), \ldots, z_i^l\left(\boldsymbol{x}^h\right)\right\}$ will have joint multivariate Gaussian distribution. The covariance is

$$K^l\left(\boldsymbol{x}, \boldsymbol{x}'\right) \equiv \mathbb{E}\left[z_i^l(\boldsymbol{x}) z_i^l\left(\boldsymbol{x}'\right)\right] = \sigma_b^2 + \sigma_w^2 \mathbb{E}\left[\phi\left(z_i^{l-1}(\boldsymbol{x})\right)\phi\left(z_i^{l-1}\left(\boldsymbol{x}'\right)\right)\right].$$

Multiple layers

Notice that $\mathbb{E}\left[\phi\left(z_i^{l-1}(\boldsymbol{x})\right)\phi\left(z_i^{l-1}(\boldsymbol{x}')\right)\right]$ is described by a mean-zero Gaussian whose covariance matrix has distinct entries $K^{l-1}(\boldsymbol{x},\boldsymbol{x}')$, $K^{l-1}(\boldsymbol{x},\boldsymbol{x})$, and $K^{l-1}(x',x')$. As such, these are the only three quantities that appear in the result. By induction, we have

$$K^l(\boldsymbol{x},\boldsymbol{x}') = \sigma_b^2 + \sigma_w^2 F_\phi\left(K^{l-1}(\boldsymbol{x},\boldsymbol{x}'), K^{l-1}(\boldsymbol{x},\boldsymbol{x}), K^{l-1}(\boldsymbol{x}',\boldsymbol{x}')\right)$$

to emphasize the recursive relationship between K^l and K^{l-1} via a deterministic function F whose form depends only on the nonlinearity ϕ. This gives an iterative series of computations which can be performed to obtain K^L for the GP describing the network's final output.

Neural network kernel

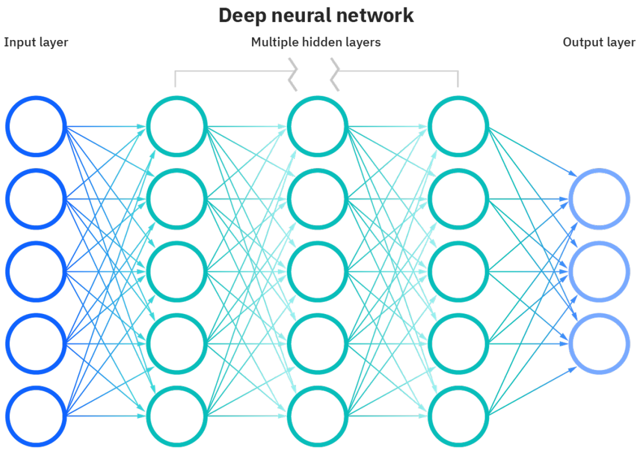


Table of Contents

Introduction

Statistical inference

Stationary covariance functions

Random covariance function

Spectral density induced covariance function

Stochastic differential equation induced covariance function

Non-stationary Covariance Functions

Differential operator induced covariance function

Neural network induced covariance function

Non-parametric covariance function

Further examples

Replicated pairs of points

We now return to the observational scheme

$$Y(t_1) = X(t_1) + \varepsilon_1,\; X(t_2) + \varepsilon_2, \cdots, Y(t_n) = X(t_n) + \varepsilon_n,$$

where ε_i's are mean-zero noises. We assume that $\mu(t) = 0$. Notice that

$$\mathbb{E} I(t_i, t_j) := \mathbb{E} Y(t_i) Y(t_j) = K(t_i, t_j) + \sigma^2 \mathbb{I}(t_i = t_j).$$

When the observed $I(t_i, t_j)$'s are sufficiently large, we can use the local polynomial or RKHS to non-parametrically estimate K.

Local linear smoother

We estimate $K(t, s)$ as γ_0 that minimizing

$$\sum_{i \neq j}^{n} \kappa_h(t_i - t, t_j - s)\left\{I(t_i, t_j) - \gamma_0 - \gamma_1(t - t_i) - \gamma_2(s - t_j)\right\}^2,$$

where κ_h is a weight function determined by the bandwidth h.

- In practice, this estimator is not realistic since n is usually not large enough to produce a large number of $I(t_i, t_j)$'s.

Replicated observed functions

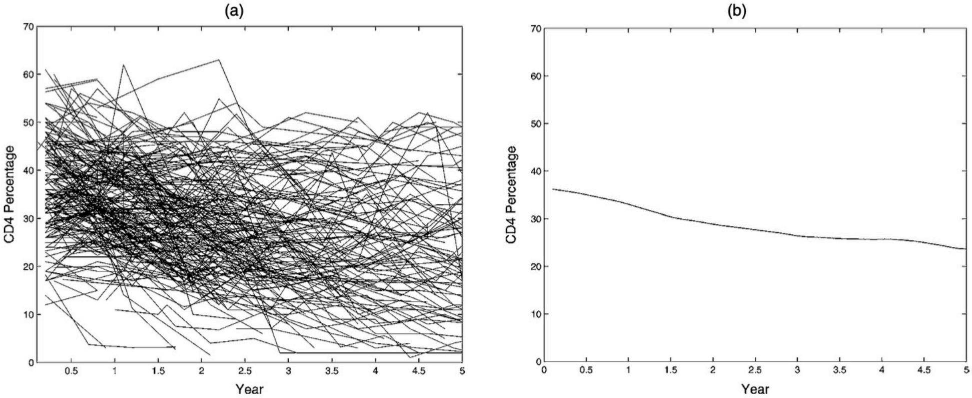
We now return to the observational scheme

$$Y_l(t_{l1}) = X_l(t_{l1}) + \varepsilon_{l1}, \cdots, Y(t_{ln_l}) = X_l(t_{ln_l}) + \varepsilon_{ln_l}$$

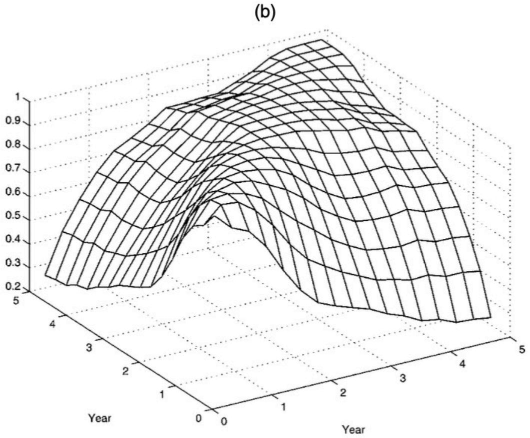
for $l = 1, \cdots, L$, where we assume that the mean functions and covariance functions are the same for different l. In this setting, we usually don't assume that $\mu(t) = 0$. When L is sufficiently large, we can non-parametrically estimate both *mu* and K under certain smoothness assumptions.

- Curse of dimensionality.

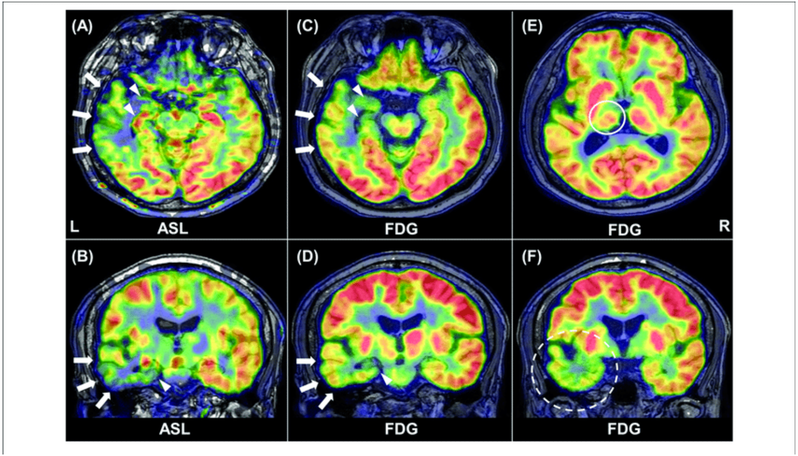
Functional data



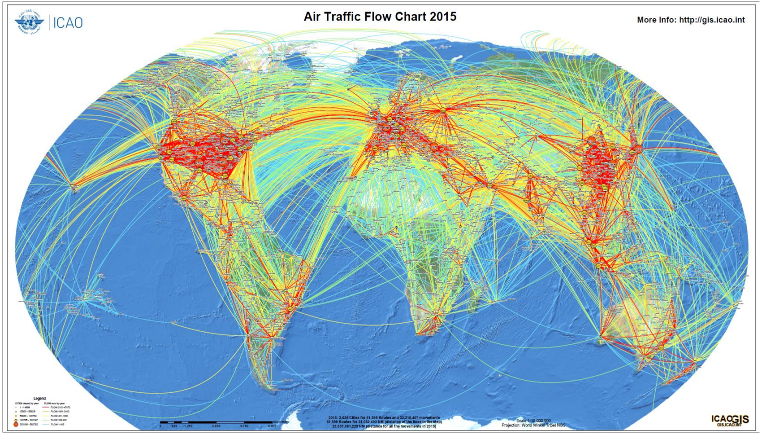
Covariance estimation



Gaussian processes on manifold



Manifold-valued Gaussian processes



Multilevel Gaussian processes