Outline

Hilbert spaces

Probabilistic models

Estimators

- Ranking based estimators
- Predictive estimators
- Indirect inference

A probabilistic toolshed

- Basic concepts
- Stochastic processes
- Partial specifications
- Gaussian processes
- Stationary stochastic processes
- Wide-sense stationarity
- Quasi-stationarity
- Frequency-domain characterization

Hilbert spaces

Let $\mathcal{V}$ be a vector space equipped with an inner product $\langle \cdot, \cdot \rangle$

1. $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$
2. $\langle \lambda u, v \rangle = \lambda \langle u, v \rangle$
3. $\langle u, v \rangle = \langle v, u \rangle^*$
4. $\langle v, v \rangle \geq 0$ with equality iff $v = 0$

Norm: $\|v\| = \sqrt{\langle v, v \rangle}$

Hilbert space $\mathcal{H}$: Complete inner product space (Cauchy sequences converge)

Extend definition to column vectors u and v of elements of $\mathcal{H}$:

$$\lfloor u, v \rfloor = M, \quad M_{i,j} = \langle u_i, v_j \rangle$$

Example 1: Consider the columns of $X \in \mathbb{R}^{N \times n_x}$ and $Y \in \mathbb{R}^{N \times n_y}$ as elements of $\mathbb{R}^N$, then

$$\lfloor X, Y \rfloor = X^T Y$$

Example 2: Let $\mathbf{x} \in \mathbb{R}^{n_x}$ and $\mathbf{y} \in \mathbb{R}^{n_y}$ be random vectors with finite second moments. Then

$$\lfloor \mathbf{x}, \mathbf{y} \rfloor = \mathbb{E}\left[\mathbf{x}\mathbf{y}^T\right]$$

Orthogonal projections

Orthogonality

An element $u \in \mathcal{H}$ is orthogonal to the subspace $\mathcal{S} \subseteq \mathcal{H}$ if

$$\langle u, v \rangle = 0 \quad \forall v \in \mathcal{S}.$$

We write $u \perp \mathcal{S}$

Projection theorem

Let $u \in \mathcal{H}$ be given and let $\mathcal{S} \subseteq \mathcal{H}$ be a closed subspace to $\mathcal{H}$. Then there exists a unique $v \in \mathcal{S}$ such that $u - v \perp \mathcal{S}$. The element v is the unique solution to

$$\min_{v \in \mathcal{S}} \|u - v\|$$

v is called the orthogonal projection of u onto $\mathcal{S}$ and is denoted $u_{\mathcal{S}}$

It follows that $u \in \mathcal{H}$ has a unique decomposition

$u = u_{\mathcal{S}} + u_{\mathcal{S}^{\perp}}$, where $u_{\mathcal{S}^{\perp}} = u - u_{\mathcal{S}} \in \mathcal{S}^{\perp}$ (subspace orthogonal to $\mathcal{S}$)

Orthogonal projections: Pythagoras relation

$$u = u_{\mathcal{S}} + u_{\mathcal{S}^\perp} \;\Rightarrow\; \|u\|^2 = \|u_{\mathcal{S}}\|^2 + \|u_{\mathcal{S}^\perp}\|^2$$

In our context often written as

$$\|u\|^2 - \|u_{\mathcal{S}}\|^2 = \|u_{\mathcal{S}^\perp}\|^2 = \|u - u_{\mathcal{S}}\|^2$$

The projection theorem:

$$\|u - v\|^2 \geq \|u - u_{\mathcal{S}}\|^2 = \|u_{\mathcal{S}^\perp}\|^2 = \|u\|^2 - \|u_{\mathcal{S}}\|^2 \geq 0 \quad \forall v \in \mathcal{S}$$

Vector version:

$$\lfloor u - v, u - v \rfloor \geq \lfloor u - u_{\mathcal{S}}, u - u_{\mathcal{S}} \rfloor = \lfloor u, u \rfloor - \lfloor u_{\mathcal{S}}, u_{\mathcal{S}} \rfloor \geq 0 \quad \forall v \in \mathcal{S}$$

Matrix inequality

Note: Projection $u_{\mathcal{S}}$ has smaller "norm" than u: $\langle u, u \rangle - \langle u_{\mathcal{S}}, u_{\mathcal{S}} \rangle \geq 0$

Orthogonal projections: Finite dimensional subspaces

Problem: Project all elements of the n_u-dimensional vector $\mathbf{u}$ on the linear span of the elements of the vector $\mathbf{y}$ (solve n_u projections simultaneously)

$$\mathcal{S} = \{\mathbf{L}\mathbf{y} : \ \mathbf{L} \in \mathbb{R}^{n_u \times n_y}\}$$

Optimality condition:

$$\begin{aligned} 0 &= \lfloor \mathbf{u} - \mathbf{L}\mathbf{y}, \mathbf{y} \rfloor = \lfloor \mathbf{u}, \mathbf{y} \rfloor - \mathbf{L} \lfloor \mathbf{y}, \mathbf{y} \rfloor \\ &\Rightarrow \ \mathbf{L}^* = \lfloor \mathbf{u}, \mathbf{y} \rfloor \lfloor \mathbf{y}, \mathbf{y} \rfloor^{-1} \\ &\Rightarrow \ \mathbf{u}_{\mathcal{S}} = \mathbf{L}^*\mathbf{y} = \lfloor \mathbf{u}, \mathbf{y} \rfloor \lfloor \mathbf{y}, \mathbf{y} \rfloor^{-1} \mathbf{y} \end{aligned}$$

Projection theorem and Pythagoras: $\mathbf{v} = \mathbf{L}\mathbf{y} \Rightarrow$

$$\lfloor \mathbf{u} - \mathbf{v}, \mathbf{u} - \mathbf{v} \rfloor \geq \lfloor \mathbf{u} - \mathbf{L}^*\mathbf{y}, \mathbf{u} - \mathbf{L}^*\mathbf{y} \rfloor = \lfloor \mathbf{u}, \mathbf{u} \rfloor - \lfloor \mathbf{u}, \mathbf{y} \rfloor \lfloor \mathbf{y}, \mathbf{y} \rfloor^{-1} \lfloor \mathbf{y}, \mathbf{u} \rfloor$$

Example: Rows of $\mathbf{U} \in \mathbb{R}^{n_u \times N}$ to be projected on the rows of $\mathbf{Y} \in \mathbb{R}^{n_y \times N}$

$$\mathbf{U}_{\mathcal{S}} = \mathbf{U}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}$$

$$0 \leq (\mathbf{U} - \mathbf{U}_{\mathcal{S}})^T(\mathbf{U} - \mathbf{U}_{\mathcal{S}}) = \mathbf{U}^T\mathbf{U} - \mathbf{U}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{U}$$

Outline

Hilbert spaces

Probabilistic models

Estimators

- Ranking based estimators
- Predictive estimators
- Indirect inference

A probabilistic toolshed

- Basic concepts
- Stochastic processes
- Partial specifications
- Gaussian processes
- Stationary stochastic processes
- Wide-sense stationarity
- Quasi-stationarity
- Frequency-domain characterization

Models and model structures

Notation:

$$\boldsymbol{\xi}^t = \begin{bmatrix} \boldsymbol{\xi}^T(0) & \dots & \boldsymbol{\xi}^T(t) \end{bmatrix}^T \in \Xi^t \subseteq \mathbb{R}^{n_{\xi^t}},\ n_{\xi^t} := \sum_{k=0}^{t} n_{\xi_t}$$

Definition

Model parameter: $\boldsymbol{\xi} = \{\boldsymbol{\xi}(t)\}_{t=0}^{\infty}$, *where* $\boldsymbol{\xi}(t) \in \boldsymbol{\Xi}(t) \subseteq \mathbb{R}^{n_{\xi_t}}$.

Model structure $\mathcal{M}(\mathbf{M}_\cdot, \boldsymbol{\Xi}) = \{\mathbf{M}_t : \boldsymbol{\Xi}^t \to \mathbb{R}^{n_z}\}_{t=1}^{\infty}$.

Model of observations: $\mathbf{z}(t) = M_t(\boldsymbol{\xi}^t),\ t = 1, 2, \dots$

Model set: $\left\{\{M_t(\boldsymbol{\xi}^t)\}_{t=1}^{\infty} :\ \boldsymbol{\xi}(t) \in \boldsymbol{\Xi}(t)\right\}$

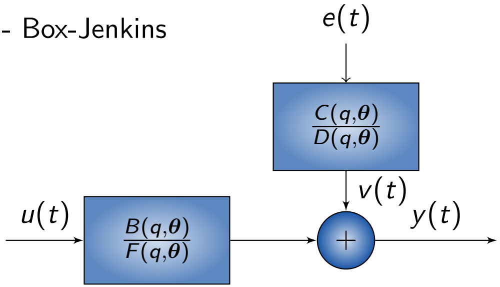
Pdf: $\{p_t : \boldsymbol{\Xi}^t \to [0, \infty)\}$ *for* $\{\boldsymbol{\xi}^t\}$

$\boldsymbol{\xi}$ *realization of* $\{p_t\}_{t=1}^{\infty} \Rightarrow \mathbf{z}(t) = M_t(\boldsymbol{\xi}^t),\ t = 1, 2, \dots$ *realization of observed signals.*

Probabilistic model structure: $\mathcal{M} = \mathcal{M}(M_\cdot, \boldsymbol{\Xi}_\cdot, p_\cdot)$

Models and model structures

LTI example - Box-Jenkins



$e(t)$
$\frac{C(q,\boldsymbol{\theta})}{D(q,\boldsymbol{\theta})}$
$v(t)$
$u(t)$
$\frac{B(q,\boldsymbol{\theta})}{F(q,\boldsymbol{\theta})}$
$+$
$y(t)$

Models and model structures

$$\mathbf{z}(t) = \begin{bmatrix} \mathbf{u}(t) \\ \mathbf{y}(t) \end{bmatrix}$$

$$\boldsymbol{\xi}(0) = \begin{bmatrix} \boldsymbol{\theta} \\ \mathrm{x}(0) \end{bmatrix}, \quad \boldsymbol{\xi}(t) = \begin{bmatrix} \overline{\mathbf{u}}(t) \\ \mathbf{e}(t) \end{bmatrix}, \quad \mathbf{x}(0) \text{ initial conditions}$$

$$\overline{y}(t) = \frac{B(q,\boldsymbol{\theta})}{F(q,\boldsymbol{\theta})}\overline{\mathbf{u}}(t) + \frac{C(q,\boldsymbol{\theta})}{D(q,\boldsymbol{\theta})}\mathbf{e}(t)$$

$$M_t(\boldsymbol{\xi}^t) = \begin{bmatrix} \overline{\mathbf{u}}^t \\ \overline{\mathbf{y}}^t \end{bmatrix}$$

$$p_t(\boldsymbol{\xi}^t) = \mathcal{N}(\mathbf{e}^t; 0, \lambda_e I)\boldsymbol{\delta}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\boldsymbol{\delta}(\overline{\mathbf{u}}^t - \tilde{\mathbf{u}}^t)\boldsymbol{\delta}(\mathbf{x}(\overline{0}) - \tilde{\mathbf{x}}(0))$$

$\boldsymbol{\theta}$, $\mathbf{x}(0)$ and $\bar{\mathbf{u}}$ deterministic.
Estimated by corresponding hyperparameters $\tilde{\boldsymbol{\theta}}$, $\tilde{\mathbf{x}}(0)$ and $\tilde{\mathbf{u}}$.
Measurement equation gives $\bar{\mathbf{u}}(t) = \mathbf{u}(t)$

Models and model structures

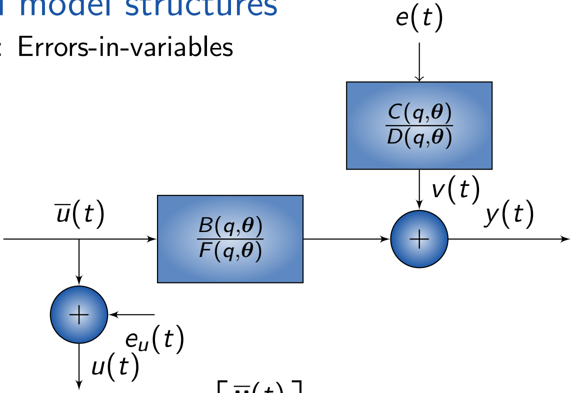
$$p_t(\boldsymbol{\xi}^t) = \mathcal{N}(\mathbf{e}^t; 0, \lambda_e I)\delta(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\delta(\overline{\mathbf{u}}^t - \tilde{\mathbf{u}}^t)\delta(\mathbf{x}(0) - \tilde{\mathbf{x}}(0))$$

$\boldsymbol{\theta}$, $\mathbf{x}(0)$ and $\bar{\mathbf{u}}$ deterministic.
Estimated by corresponding hyperparameters $\tilde{\boldsymbol{\theta}}$, $\tilde{\mathbf{x}}(0)$ and $\tilde{\mathbf{u}}$.
Consider now $\mathbf{x}(0)$ to be random $\Rightarrow$

$$p_t(\boldsymbol{\xi}^t) = \mathcal{N}(\mathbf{e}^t; 0, \lambda_e I)\delta(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\delta(\overline{\mathbf{u}}^t - \tilde{\mathbf{u}}^t)\mathcal{N}(\mathbf{x}(0), 0, \mathbf{P})$$

Models and model structures

Extension: Errors-in-variables



$e(t)$
$\frac{C(q,\boldsymbol{\theta})}{D(q,\boldsymbol{\theta})}$
$v(t)$
$\overline{u}(t)$
$\frac{B(q,\boldsymbol{\theta})}{F(q,\boldsymbol{\theta})}$
$+$
$y(t)$
$+$
$e_u(t)$
$u(t)$

$$\boldsymbol{\xi}(0) = \begin{bmatrix} \boldsymbol{\theta} \\ \mathrm{x}(0) \end{bmatrix} \quad \boldsymbol{\xi}(t) = \begin{bmatrix} \overline{\mathbf{u}}(t) \\ \mathbf{e}(t) \\ \mathbf{e}_u(t) \end{bmatrix}, \quad \mathbf{x}(0) \text{ initial conditions}$$

$$M_t(\boldsymbol{\xi}^t) = \begin{bmatrix} \overline{\mathbf{u}}^t + \mathbf{e}_u^t \\ \overline{\mathbf{y}}^t \end{bmatrix}$$

$$p_t(\boldsymbol{\xi}^t) = \mathcal{N}(\mathbf{e}^t; 0, \lambda_e I)\mathcal{N}(\mathbf{e}_u^t; 0, \lambda_u I)\boldsymbol{\delta}(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\boldsymbol{\delta}(\overline{\mathbf{u}}^t - \tilde{\mathbf{u}}^t)\boldsymbol{\delta}(\mathbf{x}(0) - \tilde{\mathbf{x}}(0))$$

$\bar{u}$ not determined exactly by measurements any longer

The set of unfalsified models

Definition

Given data $\mathbf{z}^N$*, the set of unfalsified models for the model structure* $\mathcal{M}(M_{\cdot}, p_{\cdot})$ *is defined as*

$$\mathcal{U}(\mathbf{z}^N) = \left\{ \boldsymbol{\xi} : \; M^N(\boldsymbol{\xi}^N) = \mathbf{z}^N \right\}$$

Ranking functions and pdfs

Use pdf as ranking function:

$$p_N(\boldsymbol{\xi}^N, \mathbf{z}^N) := p_N(\boldsymbol{\xi}^N) \prod_{t=1}^{N} \delta(\mathbf{z}(t) - M_t(\boldsymbol{\xi}(t)))$$

Recall that computing the average of rankings model used

$$p_N(\boldsymbol{\xi}^N | \mathbf{z}^N) := \frac{p_N(\boldsymbol{\xi}^N, \mathbf{z}^N)}{p_N(\mathbf{z}^N)}$$

This is nothing but the conditional pdf for $\boldsymbol{\xi}^N$ given observations $\mathbf{z}^N$

Marginalization: $\gamma = \gamma(\boldsymbol{\xi}^N)$

$$p_N(\gamma, \mathbf{z}^N) := \int_{\boldsymbol{\Xi}^N} p_N(\boldsymbol{\xi}^N, \mathbf{z}) \delta(\gamma - \gamma(\boldsymbol{\xi}^N)) d\boldsymbol{\xi}^N$$

Joint probability for γ and $\mathbf{z}^N$

Ranking functions and pdfs

Marginalising hyperparameter dependence

$$p_N(\mathbf{z}^N) = \int p_N(\mathbf{z}^N; \boldsymbol{\eta}) d\boldsymbol{\eta}$$

and when this quantity is finite:

$$p_N(\boldsymbol{\xi}^N, \boldsymbol{\eta} | \mathbf{z}^N) := \frac{p_N(\boldsymbol{\xi}^N, \mathbf{z}^N; \boldsymbol{\eta})}{p_N(\mathbf{z}^N)}$$

$$p_N(\boldsymbol{\eta} | \mathbf{z}^N) := \frac{p_N(\mathbf{z}^N; \boldsymbol{\eta})}{p_N(\mathbf{z}^N)}$$

Does not mean that $p_N(\boldsymbol{\xi}^N, \boldsymbol{\eta} | \mathbf{z}^N)$ and $p_N(\boldsymbol{\eta} | \mathbf{z}^N)$ should be interpreted as random

Estimators

Definition

Given a model structure $\mathcal{M}(M_{\cdot}, p_{\cdot}, \boldsymbol{\Xi}_{\cdot})$*, an estimator is a sequence of functions* $\{\hat{\boldsymbol{\xi}}^{t}\}_{t=1}^{\infty}$

$$\hat{\boldsymbol{\xi}}^{t} : \ \mathbb{R}^{n_{z_t}} \to \boldsymbol{\Xi}^{t} \subseteq \mathbb{R}^{n_{\xi_t}}$$

Outline

Hilbert spaces

Probabilistic models

Estimators

- Ranking based estimators
- Predictive estimators
- Indirect inference

A probabilistic toolshed

- Basic concepts
- Stochastic processes
- Partial specifications
- Gaussian processes
- Stationary stochastic processes
- Wide-sense stationarity
- Quasi-stationarity
- Frequency-domain characterization

Ranking based estimators

Recall maximum ranking estimator:

$$\hat{\boldsymbol{\xi}}^N(\mathbf{z}^N) = \arg\max_{\boldsymbol{\xi}^N \in \boldsymbol{\Xi}^N} p_N(\boldsymbol{\xi}^N, \mathbf{z}^N)$$

$$p_N(\boldsymbol{\xi}^N, \mathbf{z}^N) = p_N(\boldsymbol{\xi}^N | \mathbf{z}^N) p_N(\mathbf{z}^N) \;\Rightarrow\; \hat{\boldsymbol{\xi}}^N(\mathbf{z}^N) = \arg\max_{\boldsymbol{\xi}^N \in \boldsymbol{\Xi}^N} p_N(\boldsymbol{\xi}^N | \mathbf{z}^N)$$

Maximum A Posteriori (MAP) estimator $\hat{\boldsymbol{\xi}}^N_{MAP}(\mathbf{z}^N)$

Ranking based estimators

The average ranking model

$$\hat{\boldsymbol{\xi}}_A^N(\mathbf{z}^N) = \int_{\mathcal{U}(\mathbf{z}^N)} \boldsymbol{\xi}^N p_N(\boldsymbol{\xi}^N|\mathbf{z}^N) d\boldsymbol{\xi}^N = \mathbb{E}\left[\boldsymbol{\xi}^N|\mathbf{z}^N\right]$$

Posterior mean (PM) estimator $\hat{\boldsymbol{\xi}}_{PM}^N(\mathbf{z}^N)$

Ranking based estimators

Recall maximum of total ranking estimator:

$$\hat{\boldsymbol{\eta}}(\mathbf{z}^N) := \arg\max_{\boldsymbol{\eta}} p_N(\mathbf{z}^N; \boldsymbol{\eta})$$

Maximum Likelihood (ML) estimator $\hat{\boldsymbol{\eta}}_{ML}(\mathbf{z}^N)$

Actual observations have largest probability to be observed among all possible observations

PM estimator may also be used for deterministic quantities:

$$\hat{\boldsymbol{\eta}}_{PM}(\mathbf{z}^N) = \mathbb{E}\left[\boldsymbol{\eta}|\mathbf{z}^N\right] = \int \boldsymbol{\eta} p(\boldsymbol{\eta}|\mathbf{z}^N) d\boldsymbol{\eta}$$

Both model- and hyperparameters:

$$\left(\hat{\boldsymbol{\xi}}^N(\mathbf{z}^N), \hat{\boldsymbol{\eta}}(\mathbf{z}^N)\right) := \arg\max_{\boldsymbol{\xi}^N \in \boldsymbol{\Xi}^N, \boldsymbol{\eta}} p_N(\boldsymbol{\xi}^N, \mathbf{z}^N; \boldsymbol{\eta})$$

Joint MAP/ML estimator

Outline

Hilbert spaces

Probabilistic models

Estimators

- Ranking based estimators
- Predictive estimators
- Indirect inference

A probabilistic toolshed

- Basic concepts
- Stochastic processes
- Partial specifications
- Gaussian processes
- Stationary stochastic processes
- Wide-sense stationarity
- Quasi-stationarity
- Frequency-domain characterization

Predictive estimators

- Background: Probability theory $\Rightarrow$ Theory for optimal prediction of one random variable given others
- Idea: Choose model which gives best predictions
- Builds confidence in the model - not only rankings!
- Prediction essential in many applications , e.g. control, predictive maintenance and finance
- Basics:
  - Statistic: $\mathbf{s} = f(\mathbf{z}^N)$ - random under model assumption $\mathbf{s} = f(M^N(\boldsymbol{\xi}^N))$.
  - Predict: $\hat{\mathbf{s}}(\boldsymbol{\eta}) = g(\mathbf{z}^N, \boldsymbol{\eta})$
  - Minimize: $\hat{\boldsymbol{\eta}}(\mathbf{z}^N, d, f) = \arg\min_{\boldsymbol{\eta}} d(\mathbf{s}, \hat{\mathbf{s}}(\boldsymbol{\eta}))$
- Questions: What to predict ($f(\mathbf{z}^N)$) and which "distance measure" to use?
- What to predict?
  - The whole data set? Set of unfalsified models
  - ???

Predictive estimators

- What to predict and which distance measure to use?
  - $\hat{\boldsymbol{\eta}}(\mathbf{z}^N, d, f)$ random variable
  - Analyze its distribution
  - Pick d and f such that $\hat{\boldsymbol{\eta}}(\mathbf{z}^N, d, f)$ most concentrated around an $\boldsymbol{\eta}$ giving a "good" model
  - What "good" is depends on the intended model use!
    - Design variable $\boldsymbol{\rho}$
    - Optimal design $\boldsymbol{\rho}^*?\boldsymbol{\rho}^*(\boldsymbol{\xi}_o)$ ($\boldsymbol{\xi}_o$ "true" system)
    - Reward: $R(\boldsymbol{\rho}, \boldsymbol{\xi}_o)$
    - Regret: $L(\boldsymbol{\rho}, \boldsymbol{\xi}_o) = R(\boldsymbol{\rho}^*(\boldsymbol{\xi}_o), \boldsymbol{\xi}_o) - R(\boldsymbol{\rho}, \boldsymbol{\xi}_o) \geq 0$
    - Expected regret:

$$\bar{L}(\boldsymbol{\rho}^*(\hat{\boldsymbol{\xi}})) := \mathbb{E}\left[L(\boldsymbol{\rho}^*(\hat{\boldsymbol{\xi}}(\mathbf{z}), \boldsymbol{\xi})\right] = \int L(\boldsymbol{\rho}^*(\hat{\boldsymbol{\xi}}(\mathbf{M}(\boldsymbol{\xi})), \boldsymbol{\xi})p(\boldsymbol{\xi})d\boldsymbol{\xi}$$

    - With hyperparameters: $\hat{\boldsymbol{\xi}}(\mathbf{z}, \hat{\beta}(\mathbf{z}))$. Include in expectation
    - May not be optimal to use design ρ^*. Robustness considerations
    - General purpose criterion: The Mean-Square Error (MSE):

$$\mathrm{MSE}\left[\hat{\boldsymbol{\xi}}(\mathbf{z})\right] := \mathbb{E}\left[(\hat{\boldsymbol{\xi}}(\mathbf{z}) - \boldsymbol{\xi})(\hat{\boldsymbol{\xi}}(\mathbf{z}) - \boldsymbol{\xi})^T\right]$$

Indirect inference

What is the optimal estimator of a random variable $\mathbf{z}$ if no data is available?

With $\hat{\mathbf{z}}$ a constant

$$
\begin{aligned}
\mathrm{MSE}\left[\hat{\mathbf{z}}\right] &= \mathbb{E}\left[(\mathbf{z}-\hat{\mathbf{z}})(\mathbf{z}-\hat{\mathbf{z}})^T\right] \\
=&\mathbb{E}\left[(\mathbf{z}-\mathbb{E}\left[\mathbf{z}\right]+\mathbb{E}\left[\mathbf{z}\right]-\hat{\mathbf{z}})(\mathbf{z}-\mathbb{E}\left[\mathbf{z}\right]+\mathbb{E}\left[\mathbf{z}\right]-\hat{\mathbf{z}})^T\right] \\
=&\mathbb{E}\left[(\mathbf{z}-\mathbb{E}\left[\mathbf{z}\right])(\mathbf{z}-\mathbb{E}\left[\mathbf{z}\right])^T\right]+\mathbb{E}\left[(\mathbb{E}\left[\mathbf{z}\right]-\hat{\mathbf{z}})(\mathbb{E}\left[\mathbf{z}\right]-\hat{\mathbf{z}})^T\right] \\
&+\underbrace{\mathbb{E}\left[(\mathbf{z}-\mathbb{E}\left[\mathbf{z}\right])(\mathbb{E}\left[\mathbf{z}\right]-\hat{\mathbf{z}})^T\right]}_{0}+\underbrace{\mathbb{E}\left[(\mathbb{E}\left[\mathbf{z}\right]-\hat{\mathbf{z}})(\mathbf{z}-\mathbb{E}\left[\mathbf{z}\right])^T\right]}_{0} \\
=&\mathbb{E}\left[(\mathbf{z}-\mathbb{E}\left[\mathbf{z}\right])(\mathbf{z}-\mathbb{E}\left[\mathbf{z}\right])^T\right]+\mathbb{E}\left[(\mathbb{E}\left[\mathbf{z}\right]-\hat{\mathbf{z}})(\mathbb{E}\left[\mathbf{z}\right]-\hat{\mathbf{z}})^T\right] \\
\geq&\mathbb{E}\left[(\mathbf{z}-\mathbb{E}\left[\mathbf{z}\right])(\mathbf{z}-\mathbb{E}\left[\mathbf{z}\right])^T\right] = \mathrm{MSE}\left[\mathbb{E}\left[\mathbf{z}\right]\right]
\end{aligned}
$$

The mean $\mathbb{E}\left[\mathbf{z}\right]$ is the optimal estimator

Moment estimators

$$\text{Sample moments: } m_k(\mathbf{z}^N) = \frac{1}{N}\sum_{t=1}^{N} \mathbf{z}^k(t), \; k = 1, 2, \ldots$$

$$\text{Optimal estimator: } m_k(\boldsymbol{\eta}) = \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}\left[M_t^k(\boldsymbol{\xi}^t(\boldsymbol{\eta}))\right]$$

Take as many moments as dimension of $\boldsymbol{\eta}$ and solve

$$m_k(\boldsymbol{\eta}) = m_k(\mathbf{z}^N)$$

Method of moments

$$V(\boldsymbol{\eta}) = \begin{bmatrix} m_1(\mathbf{z}^N) - m_1(\boldsymbol{\eta}) \\ \vdots \\ m_K(\mathbf{z}^N) - m_K(\boldsymbol{\eta}) \end{bmatrix}^T \mathbf{W} \begin{bmatrix} m_1(\mathbf{z}^N) - m_1(\boldsymbol{\eta}) \\ \vdots \\ m_K(\mathbf{z}^N) - m_K(\boldsymbol{\eta}) \end{bmatrix}$$

$\hat{\boldsymbol{\eta}} = \arg\min_{\boldsymbol{\eta}} V(\boldsymbol{\eta})$, W corrects for different sizes of moments, e.g.

Outline

Hilbert spaces

Probabilistic models

Estimators

- Ranking based estimators
- Predictive estimators
- Indirect inference

A probabilistic toolshed

- Basic concepts
- Stochastic processes
- Partial specifications
- Gaussian processes
- Stationary stochastic processes
- Wide-sense stationarity
- Quasi-stationarity
- Frequency-domain characterization

Indirect inference

Super-simple model:

$$\mathbf{z}(t) = \mathbf{v}(t) \text{ (independent identically distributed (i.i.d.))}$$

First K moments hyperparameters: $\tilde{\boldsymbol{\eta}}_k,\ k = 1, \ldots, K$.
Estimates:

$$\hat{\tilde{\boldsymbol{\eta}}}_k(\mathbf{z}^N) = m_k(\mathbf{z})$$

Idea: If model $M(\boldsymbol{\xi}(\boldsymbol{\eta}))$ correct, data from this model should result in similar estimates for the simple model as when real data is used: For a realization of $\boldsymbol{\xi}(\boldsymbol{\eta})$

$$\hat{\tilde{\boldsymbol{\eta}}}_k(\mathbf{z}) \approx \hat{\tilde{\boldsymbol{\eta}}}_k(M(\boldsymbol{\xi}(\boldsymbol{\eta})))$$

i.e.

$$m_k(\mathbf{z}) \approx m_k(M(\boldsymbol{\xi}(\boldsymbol{\eta}))),\ k = 1, \ldots, K$$

Indirect inference

$$m_k(\mathbf{z}) \approx m_k(M(\boldsymbol{\xi}(\boldsymbol{\eta}))),\ k = 1, \ldots, K$$

But $\boldsymbol{\xi}(\boldsymbol{\eta})$ independent of data (generated by the random number generator in our computer).

Remove these by averaging:

$$m_k(\mathbf{z}) \approx \mathbb{E}\left[m_k(M(\boldsymbol{\xi}(\boldsymbol{\eta})))\right] = \frac{1}{N}\sum_{t=1}^{N} \mathbb{E}\left[M_t^k(\boldsymbol{\xi}^t(\boldsymbol{\eta}))\right] = m_k(\boldsymbol{\eta})$$

Method of moments!

What did we do?

- Intermediate model
- Estimated quantities in this model ⇒ Functions of data ($m_k(\mathbf{z})$ (statistics)
- Expected value of corresponding statistics from model matched to statistics
- Intermediate model serves to guide the choice of which statistics to use

Indirect inference

Summary:

- $\tilde{\boldsymbol{\eta}}$ hyperparameters of intermediate model
- $\hat{\tilde{\boldsymbol{\eta}}}(\mathbf{z})$ estimate
- $\boldsymbol{\eta}$ hyperparameters of model M
- $\hat{\boldsymbol{\eta}}(\mathbf{z}^N) := \arg\min_{\boldsymbol{\eta}} V_{wse}(\boldsymbol{\eta}, \mathbf{z}^N)$ where

$$V_{wse}(\boldsymbol{\eta}, \mathbf{z}) :=$$
$$\left(\hat{\tilde{\boldsymbol{\eta}}}(\mathbf{z}) - \mathbb{E}\left[\hat{\tilde{\boldsymbol{\eta}}}(M(\boldsymbol{\xi}(\boldsymbol{\eta})))\right]\right)^T W \left(\hat{\tilde{\boldsymbol{\eta}}}(\mathbf{z}) - \mathbb{E}\left[\hat{\tilde{\boldsymbol{\eta}}}(M(\boldsymbol{\xi}(\boldsymbol{\eta})))\right]\right)$$

- Different cost functions can be used, see Lecture Notes.

Prediction error methods

Idea: Predict parts of data using other parts of data

Suppose $\mathbf{z}(t) = \begin{bmatrix} \mathbf{y}^T(t) & \mathbf{u}^T(t) \end{bmatrix}^T$

$$\text{Model: } \mathbf{y}(t) = f_t(\mathbf{u}^t, \mathbf{v}^t; \boldsymbol{\theta}),\ t = 1, 2, \ldots$$

$$k\text{-step ahead predictor: } \hat{\mathbf{y}}(t+k|t;\boldsymbol{\theta}) = \hat{f}_{t+k|t}(\mathbf{u}^{t+k}, \mathbf{y}^t; \boldsymbol{\theta})$$

Prediction errors

$$\varepsilon(t+k|t;\boldsymbol{\theta}) = \mathbf{y}(t+k) - \hat{\mathbf{y}}(t+k|t;\boldsymbol{\theta}),\ t = 1, \ldots, N-k$$

Criterion (e.g.):

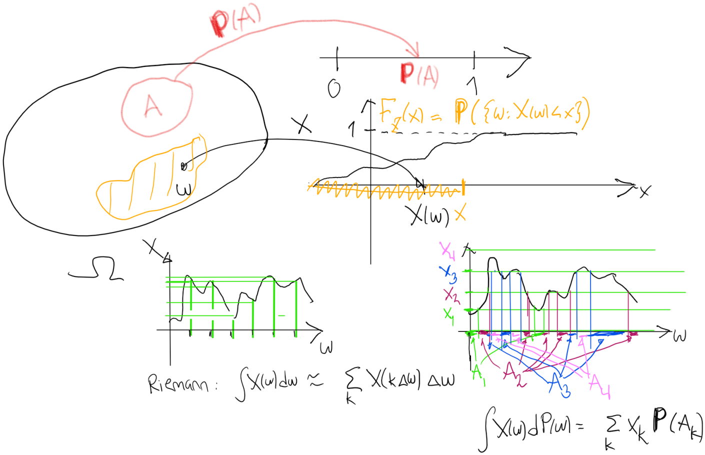
$$V_{pe,k}(\boldsymbol{\theta}, \mathbf{z}^N) := \begin{bmatrix} \varepsilon(1+k|1;\boldsymbol{\theta}) \\ \vdots \\ \varepsilon(N|N-k;\boldsymbol{\theta}) \end{bmatrix}^T W \begin{bmatrix} \varepsilon(1+k|1;\boldsymbol{\theta}) \\ \vdots \\ \varepsilon(N|N-k;\boldsymbol{\theta}) \end{bmatrix}$$

- Which $\hat{f}$ to use?
- Which criterion to use?
- ⇒ Estimation theory (next lecture)

Outline

- Hilbert spaces
- Probabilistic models
- Estimators
  - Ranking based estimators
  - Predictive estimators
  - Indirect inference
- A probabilistic toolshed
  - Basic concepts
  - Stochastic processes
  - Partial specifications
  - Gaussian processes
  - Stationary stochastic processes
  - Wide-sense stationarity
  - Quasi-stationarity
  - Frequency-domain characterization

Basic concepts



$F_X(x) = \mathbb{P}(\{\omega : X(\omega) \le x\})$

Riemann: $\int X(\omega) d\omega \approx \sum_k X(k\Delta\omega)\Delta\omega$

$$\int X(\omega) dP(\omega) = \sum_k x_k \mathbb{P}(A_k)$$

Basic concepts

- Sample space: Ω
- Probability measure: $\mathbf{P}(A)$ assigns probabilites to events A.

  i) $\mathbf{P}(\Omega) = 1$

  ii) $\mathbf{P}(\cup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \mathbf{P}(A_k)$ for disjoint events

  Not possible to assign probabilities to all sets (see ex. in LN)

  $\mathcal{F}$ set of sets for which $\mathbf{P}$ defined. Called σ-algebra

  i) $\Omega \in \mathcal{F}$

  ii) $A \in \mathcal{F} \Rightarrow A^c \in \mathcal{F}$ (complement)

  iii) $A, B \in \mathcal{F} \Rightarrow A \cup B \in \mathcal{F}$

  iv) $F_k \in \mathcal{F},\ k = 1, 2, \dots \ \Rightarrow \cup_{k=1}^{\infty} F_k \in \mathcal{F}$

iv) required to be able to compute probabilities of limits (see ex. in LN)

- Random variable: Measurable function. $\mathbf{P}(\{\omega : X(\omega) \in B)$ exists for Borel sets B
- Probability space: $(\Omega, \mathcal{F}, \mathbf{P})$

Basic concepts

- Borel set, set in $\mathcal{B}$, the Borel σ-algebra, $=$ minimal σ-algebra containing the open sets in $\mathbb{R}$.
- Probability distribution function: $\mathbf{P}_X(B) = \mathbf{P}(\{\omega : X(\omega) \in B\})$
- Distribution function: $F_X(\bar{x}) = \mathbf{P}_X(\{x : \ x \leq \bar{x}\})$

Basic concepts

Theorem

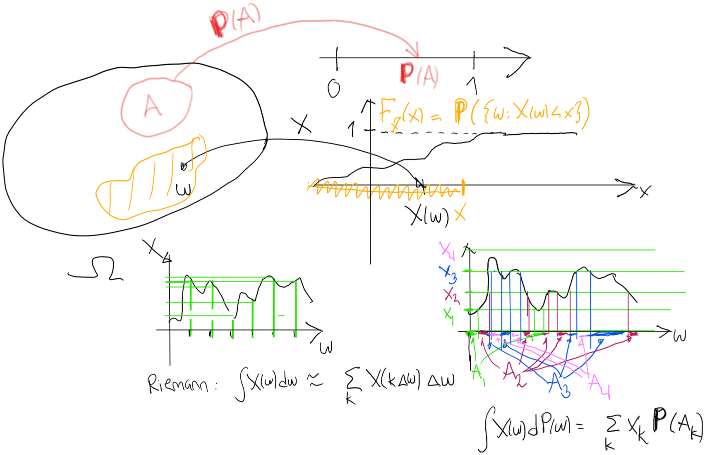
Every distribution function can be uniquely decomposed into a convex combination of a discrete, an absolutely continuous, and a continuous singular distribution function.

- Absolutely continuous: $F_X(x) = \int_{-\infty}^{x} p_X(\gamma) d\gamma$ p_X probability density function (pdf)
- Discrete: Piecewise constant. Right-continuous. At most countable number of discontinuities.
- Singular: Derivative exists almost everywhere and is zero. Continuous and can only increase on a set of measure zero.
- The distribution function can be used to compute probabilities for any Borel set.
- $\Rightarrow$ We can pretend that a r.v. is defined on $\mathbb{R}$ with probability measure F_X.

Outline

- Hilbert spaces
- Probabilistic models
- Estimators
    - Ranking based estimators
    - Predictive estimators
    - Indirect inference
- A probabilistic toolshed
    - Basic concepts
    - Stochastic processes
    - Partial specifications
    - Gaussian processes
    - Stationary stochastic processes
    - Wide-sense stationarity
    - Quasi-stationarity
    - Frequency-domain characterization

Stochastic processes



Riemann: $\int X(\omega) d\omega \approx \sum_k X(k\Delta\omega)\Delta\omega$

$$\int X(\omega) d\mathbb{P}(\omega) = \sum_k x_k \mathbb{P}(A_k)$$

Stochastic processes

Theorem (Kolmogorov)

For every set of consistent finite dimensional distributions

$$F_{t_1,\dots,t_n}(x_1,\dots,x_n) := \mathbf{P}_{\mathbf{X}}(X(t_1) \le x_1,\dots,X(t_n) \le x_n),\ t_1 < \dots < t_n$$

there exists a probability space $(\Omega, \mathcal{F}, \mathbf{P})$*, where* $\mathbf{P}$ *is unique, and a stochastic process* $\{X(t)\}$ *such that* F *is consistent with* X *and* $\mathbf{P}$*.*

Different stochastic processes can have the same distributions but different realizations

Stochastic processes

Example: (δ Kronecker's delta)
η uniformly distributed on $[0,1]$. $X(t) = \delta(t - \eta)$, $t \in [0,1]$. $\Rightarrow$

$$\mathbf{P}_{\mathbf{X}}(X(t) \in B) = \begin{cases} 1 & 0 \in B \\ 0 & \text{otherwise} \end{cases}$$

since $\eta = t$ with probability 0, and for the same reason

$$\mathbf{P}_{\mathbf{X}}(X(t_1) \in B_1, \ldots, X(t_n) \in B_n) = \begin{cases} 1 & 0 \in \cap_{k=1}^{n} B_k \\ 0 & \text{otherwise} \end{cases}$$

Let $Y(t) = 0 \cdot \eta$ for $t \in [0,1]$. $\Rightarrow$
X, Y have same finite dim. prob. dist.
However,

$$\mathbf{P}(\sup_{t \in [0,1]} Y(t) = 0) = \mathbf{P}(\sup_{t \in [0,1]} X(t) = 1) = 1$$

$\Rightarrow$ Sample paths X and Y do not coincide w.p. 1

Outline

Hilbert spaces

Probabilistic models

Estimators

- Ranking based estimators
- Predictive estimators
- Indirect inference

A probabilistic toolshed

- Basic concepts
- Stochastic processes
- Partial specifications
- Gaussian processes
- Stationary stochastic processes
- Wide-sense stationarity
- Quasi-stationarity
- Frequency-domain characterization

Partial specifications

First and second order moments

Mean function:

$$m_{\mathbf{X}}(t) := \mathbb{E}\left[\mathbf{X}(t)\right]$$

Cross-correlation function:

$$R_{\mathbf{X},\mathbf{Y}}(t,s) := \mathbb{E}\left[\mathbf{X}(t)\mathbf{Y}^T(s)\right]$$

Cross-covariance function:

$$C_{\mathbf{X},\mathbf{Y}}(t,s) := \mathbb{E}\left[(\mathbf{X}(t) - m_{\mathbf{X}}(t))(\mathbf{Y}(s) - m_{\mathbf{Y}}(s))^T\right]$$

- *Auto-correlation function* (akf): $R_{\mathbf{X},\mathbf{X}}(t,s)$
- *Covariance function*: $C_{\mathbf{X},\mathbf{X}}(t,s)$

Partial specifications

$\mathbf{X}(t)$ stochastic process with $R_{\mathbf{X},\mathbf{X}}$ as akf $\Rightarrow$

$$0 \leq \mathbb{E}\left[|\sum_i a_i^* \mathbf{X}(t_i)|^2 \right] = \sum_{i=1}^{m} \sum_{j=1}^{m} a^*(i) R_{\mathbf{X},\mathbf{X}}(t_i, t_j) a(j)$$

The opposite is true as well!

Theorem

K is a positive definite function, *i.e.*

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a^*(i) K(t_i, t_j) a(j) \geq 0, \quad \forall a(i) \in \mathbb{C}^n,\ t_i \in T,\ m \in \mathbb{N}$$

if and only if K is the akf of a stochastic process.

Modeling considerations

How do we model a family of akf's?

Obvious parametrization

$$R(t,s) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(t) \varphi_k^T(s), \quad \infty > \lambda_1 \geq \lambda_2 \geq \ldots \geq 0,$$

φ_k pre-specified basis functions, $\{\lambda_k\}$ hyperparameters

Generalization:
Let $\Phi : T \to \mathcal{H}^n$, i.e. $\Phi_i(t) \in \mathcal{H}$, $\mathcal{H}$ Hilbert space

$$R(t,s) = \lfloor \Phi(t), \Phi(s) \rfloor$$

Modeling considerations

The parametrization

$$R(t,s) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(t) \varphi_k^T(s), \quad \infty > \lambda_1 \geq \lambda_2 \geq \ldots \geq 0,$$

seems like a great idea, but maybe it does not fit the requirements for a particular application?

To study this we need to take a deviation over positive definite kernels

Positive definite kernels

T a compact domain (e.g. closed interval in $\mathbb{R}$)

Integral operators with kernel R:

$$I_R(f)(t) = \int_T R(t,s)f(s)ds$$

Maps a function f into another function. If $R \in L_\infty(T^2)$, then

$$I_R(f) : L_2(T) \to L_2(T)$$

Positive definite kernel:

$$\int_T \int_T f^*(t)R(t,s)f(s)dtds \geq 0, \quad \forall f \in L_2(T)$$

Very similar to definition of positive definite function, but not quite.
$L_2(T)$ Hilbert space $\Rightarrow$ Exists orthonormal basis $\{\boldsymbol{\varphi}_k\}$.
This basis can be chosen such that $\{\boldsymbol{\varphi}_k\}$ is bounded:
$\sup_k \sup_t |\varphi_k(t)| < \infty$

Positive definite kernels

Theorem (Mercer's theorem)

R is a bounded positive definite kernel if and only if

$$R(t,s) = \sum_{k=1}^{\infty} \lambda_k \varphi_k(t) \varphi_k^*(s),$$

where the series converges absolutely and uniformly almost everywhere, where $\lambda_k > 0$ are absolutely summable and where $\{\varphi_k\}$ is a bounded orthonormal basis for $L_2(T)$.

Positive definite functions vs kernels

There are other positive definite functions than those in Mercer's theorem. But

Theorem

Let $T = [a, b]$ be a compact interval and let $R : T \times T \to \mathbb{C}$ be continuous. Then R is a positive definite function if and only if

$$\int_T \int_T f(t)R(t,s)f(s)dtds \geq 0$$

for all complex-valued continuous functions f with domain of definition including T.

Now

- All continuous functions on $T \in L_2(T)$
- In fact they are dense in $L_2(T)$ (any function in $L_2(T)$ can be approximated arbitrarily well using a continuous function)
- $\Rightarrow$ Above can be taken as criterion for R being a positive definite kernel

Positive definite functions vs kernels

$\Rightarrow$ If we restrict $\{\boldsymbol{\varphi}_k\}$ so that R is continuous, i.e. take φ_k, $k = 1, 2, \ldots$ to be continuous, then Mercer's theorem gives:

Theorem

All continuous positive definite functions can be expressed as

$$R(t, s) = \sum_{k=1}^{\infty} \lambda_k \boldsymbol{\varphi}_k(t) \boldsymbol{\varphi}_k^*(s),$$

where $\{\boldsymbol{\varphi}_{\}}$ *is a bounded continuous orthonormal basis for* $L_2(T)$

- Complete parametrization of all continuous auto-correlation functions of a stochastic process

Outline

- Hilbert spaces
- Probabilistic models
- Estimators
  - Ranking based estimators
  - Predictive estimators
  - Indirect inference
- A probabilistic toolshed
  - Basic concepts
  - Stochastic processes
  - Partial specifications
  - Gaussian processes
  - Stationary stochastic processes
  - Wide-sense stationarity
  - Quasi-stationarity
  - Frequency-domain characterization

Gaussian processes (GP)

Pdf of a Gaussian vector:

$$\mathcal{N}(\mathbf{x}; \mathbf{m}, \mathbf{\Sigma}) := \frac{1}{\sqrt{\det 2\pi \mathbf{\Sigma}}} e^{-\frac{1}{2}(\mathbf{x}-\mathbf{m})^T \mathbf{\Sigma}^{-1} (\mathbf{x}-\mathbf{m})}$$

All finite dimensional distributions Gaussian

$$\begin{bmatrix} \mathbf{X}(t_1) \\ \vdots \\ \mathbf{X}(t_n) \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{m}(t_1) \\ \vdots \\ \mathbf{m}(t_n) \end{bmatrix}, \begin{bmatrix} C(t_1, t_1) & \dots & C(t_1, t_n) \\ \vdots & \dots & \vdots \\ C(t_n, t_1) & \dots & C(t_n, t_n) \end{bmatrix} \right), \quad \forall t_i$$

Outline

- Hilbert spaces
- Probabilistic models
- Estimators
  - Ranking based estimators
  - Predictive estimators
  - Indirect inference
- A probabilistic toolshed
  - Basic concepts
  - Stochastic processes
  - Partial specifications
  - Gaussian processes
  - Stationary stochastic processes
  - Wide-sense stationarity
  - Quasi-stationarity
  - Frequency-domain characterization

Stationary stochastic processes

Properties invariant to translation in time $\Leftrightarrow$

$$F_{t_1+\Delta,\dots,t_n+\Delta}(x_1,\dots,x_n) = F_{t_1,\dots,t_n}(x_1,\dots,x_n),$$

Consequence for first and second order statistics

$$\begin{aligned} \mathbf{m} &:= \mathbf{m}(0) = \mathbf{m}(t) \\ R_{\mathbf{X},\mathbf{Y}}(\tau) &:= R_{\mathbf{X},\mathbf{Y}}(\tau,0) = R_{\mathbf{X},\mathbf{Y}}(t,t-\tau) \\ C_{\mathbf{X},\mathbf{Y}}(\tau) &:= C_{\mathbf{X},\mathbf{Y}}(\tau,0) = C_{\mathbf{X},\mathbf{Y}}(t,t-\tau) \end{aligned}$$

Wide-sense stationarity if only the above holds

GP: Wide-sense stationarity $\Leftrightarrow {}_{(}strict) stationarity$

Outline

- Hilbert spaces
- Probabilistic models
- Estimators
  - Ranking based estimators
  - Predictive estimators
  - Indirect inference
- A probabilistic toolshed
  - Basic concepts
  - Stochastic processes
  - Partial specifications
  - Gaussian processes
  - Stationary stochastic processes
  - Wide-sense stationarity
  - Quasi-stationarity
  - Frequency-domain characterization

Wide-sense stationarity

Positivity condition

$$\sum_{i=1}^{m}\sum_{j=1}^{m} a_i^* R(t_i - t_j) a_j \geq 0, \quad \forall a_i \in \mathbb{C}^n,\ t_i \in T,\ m \in \mathbb{N}$$

$$\Leftrightarrow$$

$$\mathbf{T} = \begin{bmatrix} R(t_1 - t_1) & R(t_1 - t_2) & \dots & R(t_1 - t_m) \\ R^T(t_1 - t_2) & R(t_2 - t_2) & \dots & R(t_2 - t_m) \\ \vdots & \vdots & \ddots & \vdots \\ R^T(t_1 - t_m) & R^T(t_2 - t_m) & \dots & R(t_m - t_m) \end{bmatrix} \geq 0$$

for all Toeplitz matrices of the above type.

Outline

Hilbert spaces

Probabilistic models

Estimators

- Ranking based estimators
- Predictive estimators
- Indirect inference

A probabilistic toolshed

- Basic concepts
- Stochastic processes
- Partial specifications
- Gaussian processes
- Stationary stochastic processes
- Wide-sense stationarity
- Quasi-stationarity
- Frequency-domain characterization

Quasi-stationarity

$$\overline{\mathbb{E}}\left\{f(t)\right\} = \lim_{N\to\infty}\frac{1}{N}\sum_{t=1}^{N}\mathbb{E}\left[f(t)\right]$$

Definition

$\mathbf{X}(t)$ *is said to be a quasi-stationary signal if*

$$\begin{aligned} |m_{\mathbf{X}}(t)| &\leq C \quad \forall t \\ |R_{\mathbf{X},\mathbf{X}}(t,s)| &\leq C \quad \forall t,s \\ R_{\mathbf{X},\mathbf{X}}(\tau) :&= \overline{\mathbb{E}}\left\{\mathbf{X}(t)\mathbf{X}^T(t-\tau)\right\}, \quad \textit{exists}\ \forall\tau \end{aligned}$$

Two signals $\mathbf{X}(t)$ *and* $\mathbf{Y}(t)$ *are said to be jointly quasi-stationary if* $\left[\mathbf{X}^T(t) \quad \mathbf{Y}^T(t)\right]^T$ *is quasi-stationary.*

Outline

- Hilbert spaces
- Probabilistic models
- Estimators
  - Ranking based estimators
  - Predictive estimators
  - Indirect inference
- A probabilistic toolshed
  - Basic concepts
  - Stochastic processes
  - Partial specifications
  - Gaussian processes
  - Stationary stochastic processes
  - Wide-sense stationarity
  - Quasi-stationarity
  - Frequency-domain characterization

Frequency-domain characterization

- Recall $f(t)\overline{f(s)}$ positive definite function $\Rightarrow$
- $e^{i\omega t}e^{-i\omega s} = e^{i\omega(t-s)}$ is a positive definite function $\Rightarrow$
- $R(t) = e^{i\omega t}$ is a positive definite function $\Rightarrow$
- $R(t) = \sum_{k=1}^{n} \lambda_k e^{i\omega_k t}$ is a positive definite function
- Herglotz theorem: All positive definite functions can be generated in this way

Definition

F is a matrix valued distribution function on $[a, b]$ (or $\mathbb{R}$), if $F(a) = 0$ (or $\lim_{\omega\to-\infty} F(\omega) = 0$), F is right-continuous, $F(\omega) - F(\mu)$ is non-negative definite for all $\omega \geq \mu$.

Frequency-domain characterization

Theorem (Herglotz theorem)

$R : T \to \mathbb{R}^{m \times m}$, with $T = \mathbb{Z}$, is a positive definite function if and only if

$$R(\tau) = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{i\omega\tau} dF(\omega)$$

where F is an $m \times m$ matrix valued distribution function on $[-\pi, \pi]$, called the spectral distribution function

Bochner's theorem in continuous time, see LN

Frequency-domain characterization

Corollary

Suppose that $R : T \to \mathbb{R}^{m\times m}$, with $T = \mathbb{Z}$, is absolutely summable $\sum_{\tau=-\infty}^{\infty} \|R(\tau)\|_F < \infty$. Then R is a positive definite function if and only if

$$R(\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i\omega\tau} Q(\omega) d\omega$$

for some continuous function $Q \in L_1(\mathbb{R})$, satisfying $Q(\omega) \geq 0$, $\omega \in [-\pi, \pi]$, called the spectrum.

Notation $\Phi(e^{i\omega}) = Q(\omega)$

$$\mathbb{E}\left[\mathbf{X}(t)\mathbf{X}^T(t)\right] = R_{\mathbf{X}}(0) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \Phi_{\mathbf{x}}(e^{i\omega}) d\omega$$

$\Phi_{\mathbf{X}}$ distribution of signal power over frequencies

Modeling considerations

Some parametrizations of spectra:

- pdf's (i.e. use characteristic functions as akf's)
- $\Phi(\omega) = \sum_{k=1}^{\infty} \alpha_k \mathcal{B}_k(e^{i\omega}),\ \mathcal{B}_k(e^{i\omega}) \geq 0,\ \alpha_k \geq 0,\ k = 1, 2, \ldots$
- $\Phi(e^{i\omega}) = \tilde{H}(e^{i\omega})\tilde{H}^*(e^{i\omega})$

The last can be given a filtering interpretation:

G BIBO stable

$$y(t) = G(q)u(t) \quad \Rightarrow \quad \Phi_{yy}(e^{i\omega}) = G(e^{i\omega})\Phi_{uu}(e^{i\omega})G^*(e^{i\omega})$$

More on this in the next lecture

A swatch of building blocks

Let us study a GP

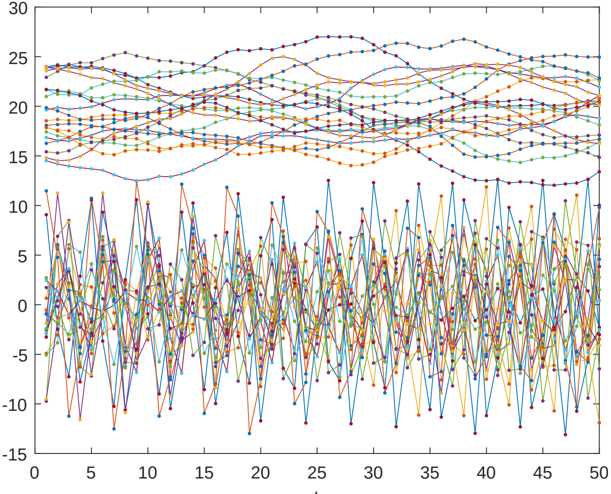
$$f(\cdot) \sim \mathcal{N}(0, K(\cdot, \cdot))$$

for different choices of kernel K

A swatch of building blocks

Disturbances and noise: Behaviour often does not change over time

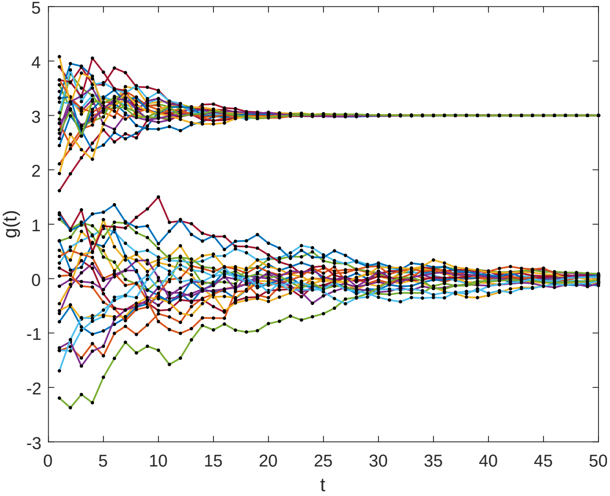
$$K(v(t), v(s)) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |H(e^{i\omega}|^2 e^{i\omega(t-s)} d\omega$$



A swatch of building blocks

Impulse responses of stable linear systems: Decays exponentially

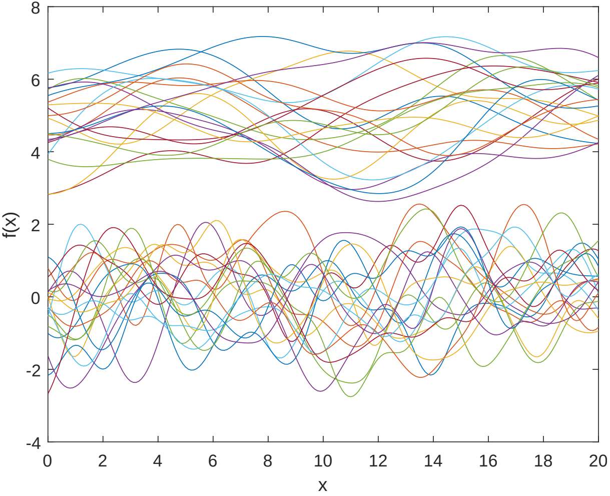
$$K(g(t), g(s)) = \eta_1 \, \eta_2^{\min(t,s)}, \quad |\eta_2| < 1$$



A swatch of building blocks

Gaussian kernel: Often used when modeling a non-linear function

$$K(f(x), f(y)) = \eta_1 \; e^{-\frac{|x-y|^2}{2\eta_2}}, \; \eta_1 > 0, \; \eta_2 > 0$$



Summary

- Hilbert spaces
- Probabilistic models
- Estimators
  - Ranking based estimators
  - Predictive estimators
  - Indirect inference
- A probabilistic toolshed
  - Basic concepts
  - Stochastic processes
  - Partial specifications
  - Gaussian processes
  - Stationary stochastic processes
  - Wide-sense stationarity
  - Quasi-stationarity
  - Frequency-domain characterization