Data Driven Modeling

Lecture 2

Håkan Hjalmarsson

KTH - Royal Institute of Technology

hjalmars@kth.se

May 19, 2020

Outline

Hilbert spaces

Let $\mathcal{V}$ be an inner product space, i.e. 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: $\|v\| = \sqrt{\langle v, v \rangle}$

Complete space: Cauchy sequences converge

$$x_n \in \mathcal{H}, \lim_{m,n \to \infty} \|x_n - x_m\| \to 0 \;\Leftrightarrow\; x \in \mathcal{H} : \lim_{n \to \infty} \|x_n - x\| = 0$$

A Hilbert space is a complete inner product space

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

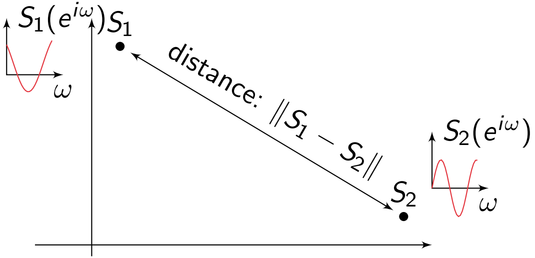
$$\lfloor X, Y \rfloor = XY^T$$

$L_2(\mathbb{T})$ - an example of a non-trivial Hilbert space

$$\text{Inner product: } \langle S, V \rangle = \frac{1}{2\pi} \int_{-\pi}^{\pi} \mathrm{Trace}\left\{ V^*(e^{i\omega}) S(e^{i\omega}) \right\} d\omega$$

$$L_2(\mathbb{T}) = \left\{ S : \ \|S\|_2^2 := \langle S, S \rangle = \|S\|^2 < \infty \right\}$$

Recall $L_2(\mathbb{T})$ consists of equivalence classes:



Functions grouped together that satisfies

$$0 = \|S_1 - S_2\|_2^2 = \frac{1}{2\pi} \int_{-\pi}^{\pi} |S_1(e^{i\omega}) - S_2(e^{i\omega})|^2 d\omega$$

Orthogonal projections

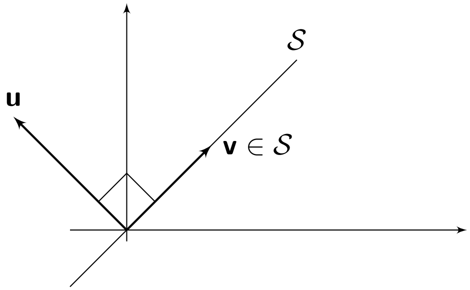
Inner product provides a geometry:

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 $u \perp \mathcal{S}$



Orthogonal projections

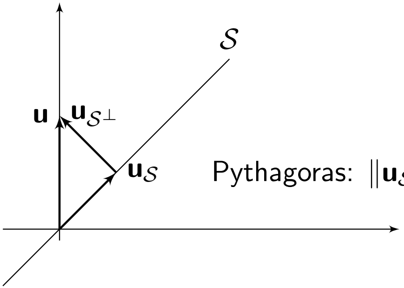
Projection theorem

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:



Pythagoras: $\|\mathbf{u}_{\mathcal{S}}\|^2 + \|\mathbf{u}_{\mathcal{S}^\perp}\|^2 = \|\mathbf{u}\|^2$

Orthogonal projections: Pythagoras relation

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:

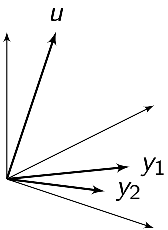
$$\lfloor u - v, u - v \rfloor \geq \lfloor u - u_{\mathcal{S}}, u - u_{\mathcal{S}} \rfloor = \lfloor u, u \rfloor - \lfloor u_{\mathcal{S}}, u_{\mathcal{S}} \rfloor \geq 0 \quad \forall v \in \mathcal{S}$$

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 $u \in \mathcal{H}$ on the linear span of elements in $\mathbf{y}$



$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$\mathcal{S} = \{\mathbf{ly} : \mathbf{l} \in \mathbb{R}^{1 \times n_y}\}$$

Let $\mathbf{ly}$ be candidate for the projection.
Try to find $\mathbf{l}$ s.t.: $0 = \langle u - \mathbf{ly}, y_k \rangle$, $k = 1, \ldots, n_y$.
Compact form:

$$0 = \lfloor u - \mathbf{ly}, \mathbf{y} \rfloor = \lfloor u, \mathbf{y} \rfloor - \mathbf{l} \lfloor \mathbf{y}, \mathbf{y} \rfloor \Rightarrow \mathbf{l}^* = \lfloor u, \mathbf{y} \rfloor \lfloor \mathbf{y}, \mathbf{y} \rfloor^{-1}$$

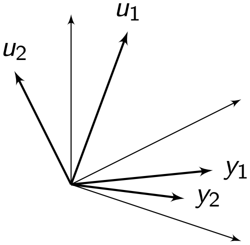
$$\Rightarrow u_{\mathcal{S}} = \mathbf{l}^* \mathbf{y} = \lfloor u, \mathbf{y} \rfloor \lfloor \mathbf{y}, \mathbf{y} \rfloor^{-1} \mathbf{y}$$

Projection theorem and Pythagoras: $v \in \mathcal{S}$ i.e. $v = \mathbf{ly} \Rightarrow$

$$\begin{aligned} \lfloor u - v, u - v \rfloor \geq \lfloor u - \mathbf{l}^* \mathbf{y}, u - \mathbf{l}^* \mathbf{y} \rfloor &= \lfloor u, u \rfloor - \lfloor \mathbf{l}^* \mathbf{y}, \mathbf{l}^* \mathbf{y} \rfloor \\ &= \lfloor u, u \rfloor - \lfloor u, \mathbf{y} \rfloor \lfloor \mathbf{y}, \mathbf{y} \rfloor^{-1} \lfloor \mathbf{y}, u \rfloor \end{aligned}$$

Orthogonal projections: Finite dimensional subspaces

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



$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

$$\mathcal{S} = \{\mathbf{l}\mathbf{y} : \ \mathbf{l} \in \mathbb{R}^{1 \times n_y}\}$$

Let projections be $\mathbf{L}\mathbf{y}$: $\mathbf{L} \in \mathbb{R}^{n_u \times n_y}$

$$\text{Same formulas: } \mathbf{u}_{\mathcal{S}} = \mathbf{L}^* \mathbf{y} = \lfloor \mathbf{u}, \mathbf{y} \rfloor \lfloor \mathbf{y}, \mathbf{y} \rfloor^{-1} \mathbf{y}$$

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: Project rows of $\mathbf{U} \in \mathbb{R}^{n_u \times N}$ on rows of $\mathbf{Y} \in \mathbb{R}^{n_y \times N}$

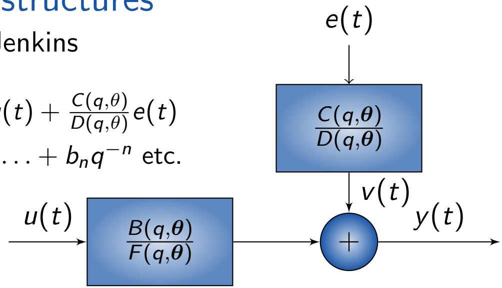
$$\mathbf{U}_{\mathcal{S}} = \mathbf{U}\mathbf{Y}^T(\mathbf{Y}\mathbf{Y}^T)^{-1}\mathbf{Y}, \ (\mathbf{U} - \mathbf{U}_{\mathcal{S}})^T(\mathbf{U} - \mathbf{U}_{\mathcal{S}}) = \mathbf{U}^T\mathbf{U} - \mathbf{U}^T\mathbf{Y}(\mathbf{Y}^T\mathbf{Y})^{-1}\mathbf{Y}^T\mathbf{U}$$

Outline

Models and model structures

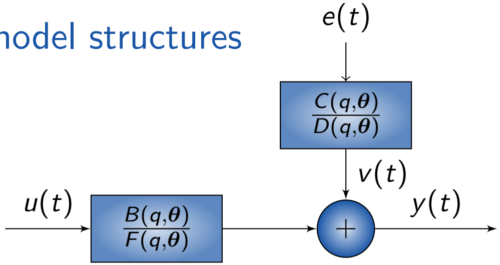
LTI example - Box-Jenkins

$$y(t) = \frac{B(q,\theta)}{F(q,\theta)} u(t) + \frac{C(q,\theta)}{D(q,\theta)} e(t)$$
$$B(q) = b_1^{-1} + \ldots + b_n q^{-n} \text{ etc.}$$



- Observations: $\mathbf{z}$
- Model parameters: $\boldsymbol{\xi} \in \boldsymbol{\Xi}$, everything that is unknown
- Model structure: Map M from model par. to observations
- Model of observations: $M(\boldsymbol{\xi})$
- Model set: All models of observations $\{M(\boldsymbol{\xi}) : \boldsymbol{\xi} \in \boldsymbol{\Xi}\}$
- Model parameter distribution: Pdf for model parameters $p(\boldsymbol{\xi})$
- Need to account for that a model is dynamic and arbitrary number of observation
- Notation: $\mathbf{x}^t = \left[\mathbf{x}(0)^T \quad \ldots \mathbf{x}(t)^T\right]^T$

Models and model structures



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

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

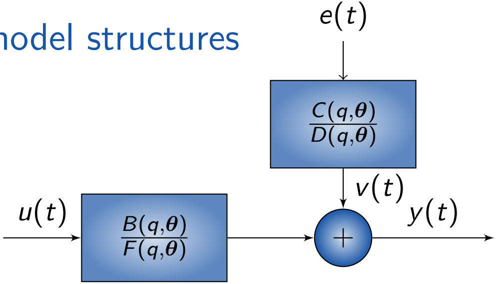
$$\overline{y}(t) = \frac{B(q,\boldsymbol{\theta})}{F(q,\boldsymbol{\theta})}\overline{u}(t) + \frac{C(q,\boldsymbol{\theta})}{D(q,\boldsymbol{\theta})}e(t)$$

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

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

Hyperparameters: $\boldsymbol{\eta}^t = \begin{bmatrix} \lambda_e & \tilde{\boldsymbol{\theta}}^T & \tilde{\mathbf{x}}^T(0) & (\tilde{\mathbf{u}}^t)^T \end{bmatrix}^T$

Models and model structures



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

Hyperparameters: $\boldsymbol{\eta}^t = \begin{bmatrix} \lambda_e & \tilde{\boldsymbol{\theta}}^T & \tilde{\mathbf{x}}^T(0) & (\tilde{\mathbf{u}}^t)^T \end{bmatrix}^T$

- All model parameters included in the probabilistic description
- Use Dirac-functions for deterministic parameters
- Hyperparameters:
  - Parameters not needed to generate the model response
  - Used as dummy variables for deterministic model parameters
  - Split between model- and hyperparameters not unique

Consider now $\mathbf{x}(0)$ to be random $\Rightarrow$

$$p_t(\boldsymbol{\xi}^t; \boldsymbol{\eta}^t) = \mathcal{N}(\mathbf{e}^t; 0, \lambda_e I)\delta(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\delta(\overline{\mathbf{u}}^t - \tilde{\mathbf{u}}^t)\mathcal{N}(\mathbf{x}(0), 0, 10I)$$

Models and model structures

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: $\mathbf{z}(t) = M_t(\boldsymbol{\xi}^t),\ t = 1, 2, \ldots$

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

Model parameter distribution: $p = \{p_t : \boldsymbol{\Xi}^t \to [0, \infty)\}$ *for* $\{\boldsymbol{\xi}^t\}$

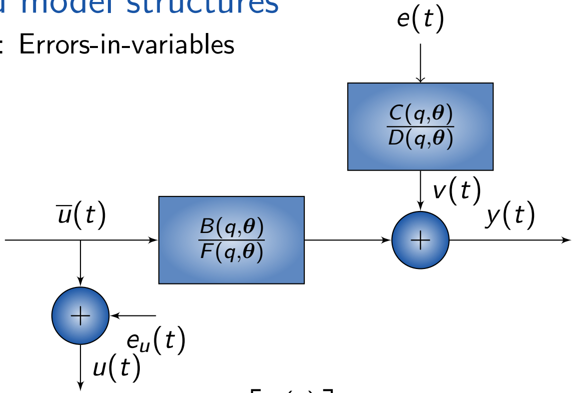
$\boldsymbol{\xi}$ *realization of* $\{p_t\}_{t=1}^{\infty} \Rightarrow M_t(\boldsymbol{\xi}^t),\ t = 1, 2, \ldots$ *realization of model.*

Hyperparameters: Parametrization $\boldsymbol{\eta}$ *of* p

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

Models and model structures

Extension: Errors-in-variables



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

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

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

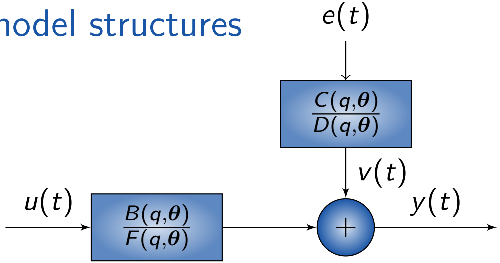
The set of unfalsified models

Definition

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

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

Models and model structures



$$
\begin{aligned}
\mathbf{z}(t) &= \begin{bmatrix} u(t) & y(t) \end{bmatrix}^T \\
\overline{y}(t) &= \frac{B(q,\boldsymbol{\theta})}{F(q,\boldsymbol{\theta})}\overline{u}(t) + \frac{C(q,\boldsymbol{\theta})}{D(q,\boldsymbol{\theta})}e(t) \\
M_t(\boldsymbol{\xi}^t) &= \begin{bmatrix} \overline{u}(t) & \overline{y}(t) \end{bmatrix}^T \\
p_t(\boldsymbol{\xi}^t; \boldsymbol{\eta}^t) &= \mathcal{N}(\mathbf{e}^t; 0, \lambda_e I)\delta(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})\delta(\overline{\mathbf{u}}^t - \tilde{\mathbf{u}}^t)\delta(\mathbf{x}(0) - \tilde{\mathbf{x}}(0))
\end{aligned}
$$

Hyperparameters: $\boldsymbol{\eta}^t = \begin{bmatrix} \lambda_e & \tilde{\boldsymbol{\theta}}^T & \tilde{\mathbf{x}}^T(0) & (\tilde{\mathbf{u}}^t)^T \end{bmatrix}^T$

$z(t) = M_t(\boldsymbol{\xi}^t) \Rightarrow \bar{u}(t) = u(t) \Rightarrow \tilde{u}(t) = u(t)$

Ranking functions and the probability distribution

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

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

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

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

Outline

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 hyperparameter 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:

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

Combinations $\Rightarrow$ Many variations possible

Outline

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

Predictive estimators

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

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

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

The mean $\mathbb{E}[\mathbf{z}]$ 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.

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!
  - General purpose criterion: The Mean-Square Error (MSE):

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

and its matrix version

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

and the equivalent for hyperparameter estimators

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?
- $\Rightarrow$ Estimation theory (next lecture)

Outline

Indirect inference

Super-simple model:

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

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

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

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 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 $\Rightarrow$ 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

Indirect inference

Generalization:

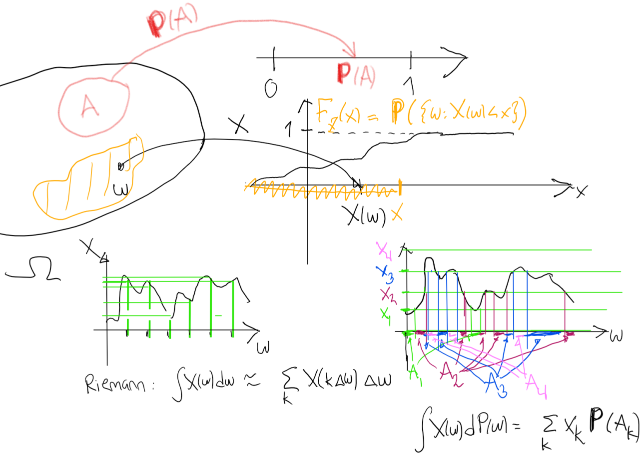
- $\tilde{\boldsymbol{\eta}}$ hyperparameters of intermediate model
- $\hat{\tilde{\boldsymbol{\eta}}}(\mathbf{z})$ estimate
- $\boldsymbol{\eta}$ hyperparameters of model M
- $\hat{\boldsymbol{\eta}}(\mathbf{z}^N) := \arg\min_{\boldsymbol{\eta}} V_{wse}(\boldsymbol{\eta}, \mathbf{z}^N)$ where

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

- Different cost functions can be used, see LN.

Outline

Basic concepts



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

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

- Probability space: $(\Omega, \mathcal{F}, \mathbf{P})$

Basic concepts

- Borel σ-algebra: minimal σ-algebra containing the open sets in $\mathbb{R}$.
- Random variable: Measurable function, i.e. $\mathbf{P}(\{\omega : X(\omega) \in B)$ exists for all Borel sets B
- Probability distribution function: $\mathbf{P}_X(B) = \mathbf{P}(\{\omega : X(\omega) \in B\})$
- Distribution function: $F_X(\bar{x}) = \mathbf{P}_X(\{x : \ x \leq \bar{x}\})$

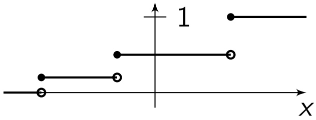
Basic concepts

Theorem

Every distribution function F can be uniquely decomposed into

$$F(x) = \alpha F_a(x) + \beta F_d(x) + \gamma F_s(x),\ \alpha,\beta,\gamma \geq 0,\ \alpha+\beta+\gamma = 1$$

- F_a absolutely continuous: $F_a(x) = \int_{-\infty}^{x} p_X(\gamma) d\gamma$, p_X probability density function (pdf)
- F_d discrete: Piecewise constant. Right-continuous. At most countable number of discontinuities.



- F_s 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. $(\mathbb{R}, \mathcal{B}, "F")$ probability space

Outline

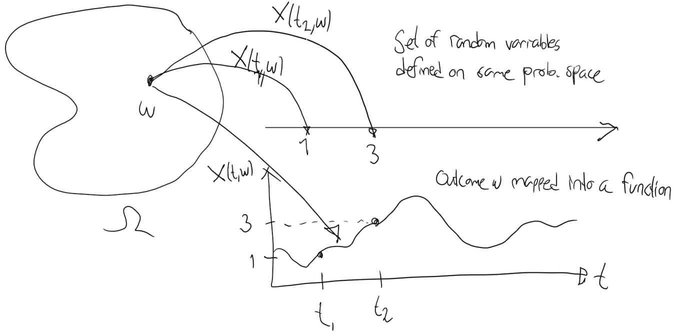
Hilbert spaces

Probabilistic models

Estimators
- Ranking based estimators
- Predictive estimators
- Indirect inference

A probabilistic toolshed
- Basic concepts
- Stochastic processes
- Partial specifications
- Gaussian processes

Stochastic processes



Stochastic processes

Theorem (Kolmogorov)

For every set of consistent finite dimensional distributions

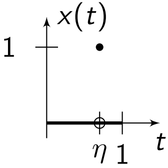
$$F_{t_1,\ldots,t_n}(x_1,\ldots,x_n) := \mathbf{P_X}(X(t_1) \le x_1, \ldots, X(t_n) \le x_n),\ t_1 < \ldots < t_n$$

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: η uniformly distributed on $[0, 1]$.



$$\mathbf{P}_x(x(t) = 1) = \mathbf{P}(\eta = t) = 0 \;\Rightarrow$$

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

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

Let $y(t) = 0 \cdot \eta$ for $t \in [0, 1]$. $\Rightarrow$ x & y have same finite dim. 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 of x and y do not coincide w.p. 1

Outline

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 $R \in L_\infty(T^2)$[1], 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 $\{\varphi_k\}$.
Can be chosen s.t. $\{\varphi_k\}$ is bounded: $\sup_k \sup_t |\varphi_k(t)| < \infty$

[1] T^2 is shorthand for $T \times T$

Positive definite kernels

Theorem (Mercer's theorem)

T compact domain. 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 $\{\varphi_k\}$ so that R is continuous, i.e. take φ_k, $k = 1, 2, \dots$ to be continuous, then Mercer's theorem gives:

Theorem

T finite interval. All continuous positive definite functions can be expressed as

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

where $\{\varphi_{\}}$ is a bounded continuous orthonormal basis for $L_2(T)$

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

Outline

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

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