MGD PROPRIETIES

7/14/2023

$$x \in \mathbb{R}^d, \mu \in \mathbb{R}^d, \Sigma \in S^d \Rightarrow x \sim N(\mu, \Sigma)$$

$$\boxed{P(x;\mu,\Sigma) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right\}}$$

WHERE OUR PARAMETERS ARE DEFINED AS FOLLOWS (EXAMPLE IN 2 DIMENSIONS)

$$x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_A \\ \mu_B \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{AA} & \Sigma_{Ab} \\ \Sigma_{BA} & \Sigma_{BB} \end{bmatrix}$$

1. NORMALIZATION

$$\boxed{\int_x P(x;\mu,\Sigma) = 1}$$

2. MARGINALIZATION

$$P(x_{\underset{B}{A}}) = \int_{x_{\underset{A}{B}}} P(x_{\underset{B}{A}}, x_{\underset{A}{B}} ; \mu, \Sigma)\, dx_B$$

$$x_A \sim N(\mu_{\underset{B}{A}}, \Sigma_{\underset{BB}{AA}})$$

WE OBTAIN A MARGINALIZED DISTRIBUTION (WHICH IS ALSO A NGD)

ex

$$x = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_d \end{bmatrix} \quad \Sigma = \begin{bmatrix} \Sigma_{11} & & \\ & \ddots & \\ & & \Sigma_{dd} \end{bmatrix} \xrightarrow{\text{MARGINALIZED}}$$

- $x \sim N(\mu, \Sigma)$
- $x_1 \sim N(\mu_1, \Sigma_{11})$
- $\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & & \\ & \ddots & \\ & & \Sigma_{33} \end{bmatrix} \right)$

3. CONDITIONING

$$P(x_A | x_B) = \frac{P(x_A, x_B ; \mu, \Sigma)}{\int_{x_A} P(x_A, x_B ; \mu, \Sigma)} = \frac{P(x_A, x_B ; \mu, \Sigma)}{P(x_B ; \mu, \Sigma)}$$

WE ARE NOW ABLE TO DEFINE A NEW NGD

$$x_A | x_B \sim N\Big(\mu_A + \underbrace{\Sigma_{AB}\Sigma_{BB}^{-1}}_{\frac{\rho\sigma_1\sigma_2}{\sigma_2^2}}(x_B - \mu_B),\ \Sigma_{AA} - \underbrace{\Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}}_{\frac{\rho\sigma_1\cancel{\sigma_2}}{\cancel{\sigma_2^2}} \cdot \rho\cancel{\sigma_2}\sigma_1 = \rho^2\sigma_1^2}\Big) \quad (*)$$

ex

LET'S CONSIDER TWO DIMENSIONS

$$x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \quad \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

$$\Sigma = \begin{pmatrix} \overset{\Sigma_{11}}{\sigma_1^2} & \overset{\Sigma_{12}}{\rho\sigma_1\sigma_2} \\ \overset{\Sigma_{21}}{\rho\sigma_1\sigma_2} & \overset{\Sigma_{22}}{\sigma_2^2} \end{pmatrix}$$

WHERE σ IS THE STD. DEVIATION

σ^2 THE VARIANCE

ρ THE CORRELATION TERM

THEN, FROM (*) WE GET

$$\boxed{x_1 | x_2 \sim N\left(\mu_1 + \rho\sigma_1 \frac{(x_2 - \mu_2)}{\sigma_2},\ \sigma_1^2 - \rho^2\sigma_1^2\right)}$$

4. SUMMATION

SUPPOSE TWO INDIPENDENT RANDOM VARIABLES

$$\begin{cases} y \sim N(\mu, \Sigma) \\ y' \sim N(\mu', \Sigma') \end{cases} \longrightarrow \boxed{y + y' \sim N(\mu + \mu', \Sigma + \Sigma')}$$

BAYESIAN LINEAR REGRESSION

$$S = \{x^{(i)}, y^{(i)}\}_{i=n}^{n} \quad , \quad y^{(i)} = \theta^T x^{(i)} + \varepsilon^{(i)} \quad i = 0, \dots, n \quad \text{AND} \quad \varepsilon^{(i)} \sim N(0, \sigma^2)$$

LET'S NOW ADD THE BAYESIAN ASSUMPTION

$$\boxed{\theta \sim N(0, \tau^2 I)}$$

STEPS OF BAYESIAN METHOT

PRIOR DISTRIB.

POSTERIOR

NORMAL DISTR.

$$P(\theta) \longrightarrow \text{OBSERVE } S \longrightarrow P(\theta | S) \overset{B}{=} \frac{P(\theta) \prod_{i=1}^{n} \overbrace{P(y^{(i)} | x^{(i)}, \theta)}}{\int_\theta P(\theta) \prod_{i=1}^{n} P(y^{(i)} | x^{(i)}, \theta) d\theta}$$

AND WE KNOW THAT

$$\theta | S \sim N\left(\frac{1}{\sigma^2} A^{-1} X^T \vec{y}, A^{-1}\right) \quad \text{WHERE} \quad A = \frac{1}{\sigma^2} X^T X \cdot \frac{1}{\tau^2} I$$

$$\theta | S \sim N\left(\frac{1}{\sigma^2}\left(\frac{1}{\sigma^2} X^T X + \frac{1}{\tau^2} I\right)^{-1} X^T \vec{y}, A^{-1}\right) = \boxed{N\left(\left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1} X^T \vec{y}, A^{-1}\right) \sim \theta | S}$$

GIVEN THE POSTERIOR PREDICTIVE DISTRIBUTION, WE CAN ESTIMATE $\hat{\Theta}$ AS FOLLOWS

$$\boxed{\hat{\Theta} = (X^T X)^{-1} X^T \vec{y}}$$

AND DEFINE ITS DISTRIBUTION AS

$$\boxed{\Theta \sim N\left(\left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1} X^T \vec{y}, \left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1} \right)}$$

AND THE PREDICTION $\hat{y}$ WILL BE AS FOLLOWS

$$\boxed{\underset{\text{PRED}}{\hat{y}} = \underset{\text{EST. }\theta}{\hat{\Theta}}^T \underset{\text{NEW EX}}{x_*}}$$

GAUSSIAN PROCESS

LET'S MAKE THIS ASSUMPTION

VARIABLES, SCALARS	$\longrightarrow$	VECTOR "COLLECTION OF SCALARS" $(x_1, \dots, x_d)$	$\longrightarrow$	FUNCTIONS! = VECTOR W/ ARBITRARY INDEXES (INDEX SET: $\mathbb{D}, \mathbb{I}$)

IF WE APPLY TO A MNGD A FUNCTION INSTEAD OF VECTORS, WE GET A GAUSSIAN PROCESS

$$\overset{\mathbb{R}^d}{\overset{\cup\!\!\mid}{}} \quad x \sim \mathcal{N}(\mu, \Sigma) \quad \text{VS} \quad \vec{f} \sim \mathcal{N}(\mu, \Sigma) \quad \overset{\mathbb{R}^d}{\overset{\cup\!\!\mid}{}}$$

$$f_0 = f(0) \sim \mathcal{N}(\mu_0, \Sigma_{00})$$

$$\vdots \quad \infty \quad \vdots$$

$$f_d = f(d) \sim \mathcal{N}(\mu_d, \Sigma_{dd})$$

WE HAVE A TRAINING SET $\{\underbrace{x^{(i)}}_{\in \mathbb{R}^d}, y^{(i)}\}_{i=1}^n$, NOT DEFINING A MGD OVER $x^{(i)}$'S. $x^{(i)}$'S WILL BE THE INDEXES

$x^{(i)} \in \mathcal{X}$ ABSTRACT SPACE ,

$$\begin{array}{c} f_{x^{(1)}} \\ f_{x^{(2)}} \\ f_{x^{(3)}} \\ \\ \\ \end{array} \begin{bmatrix} f(x^{(1)}) \\ f(x^{(2)}) \\ f(x^{(3)}) \\ \vdots \\ f(x^{(n)}) \end{bmatrix} = \begin{bmatrix} f_{x^{(1)}} \\ f_{x^{(2)}} \\ f_{x^{(3)}} \\ \vdots \\ f_{x^{(n)}} \end{bmatrix}$$

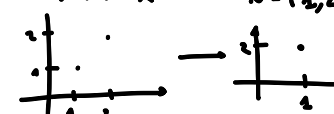
THIS ENTITY INDEXED W/ ANYTHING $\in \mathcal{X}$ (∞ DIMENSIONAL)

$$f_{\mathcal{X}}(\cdot) \sim GP(m(\cdot), K(\cdot, \cdot)) \quad \mathbb{D}: x^{(1)}, \dots, x^{(n)}; \ \mathcal{X} \mapsto \mathbb{R}^d$$

$$z \sim \mathcal{N}(\mu, \Sigma) \in \mathbb{R}^d \quad \mathbb{D}: 1, \dots, d$$

INDEX: $1, \dots, d$ (0 DIMENSIONAL)

$f(x) = x$ $\qquad$ $\mathcal{N} = (1, 2)$



HOW LIKELY A DETERMINISTIC FUNCTION IS

GP → DISTRIBUTION OVER FUNCTIONS, A SAMPLE IS A FUNCTION

N → DISTRIBUTION OVER VECTORS, A SAMPLE IS A VECTOR

$$z : [1, d] \longmapsto \mathbb{R}, \quad \mu : [1, d] \longmapsto \mathbb{R}, \quad \overset{\text{PSD}}{\Sigma} : [1, d] \times [1, d] \longmapsto \mathbb{R}$$

KERNEL (COVARIANCE FUNCTIONS)

$$f : \mathcal{X} \longmapsto \mathbb{R}, \quad \mu : \mathcal{X} \longmapsto \mathbb{R}, \quad K : \mathcal{X} \times \mathcal{X} \longmapsto \mathbb{R}$$

$$\boxed{\begin{gathered} z^T M \geq \ldots\ldots \geq 0 \\ \downarrow \\ \int_z \int_x f(z) K(z, x) f(x) \, dx \, dz \end{gathered}}$$

VS

$$\boxed{\begin{gathered} a^T b \\ \sum_i a_i b_i \\ \int_z a(z) b(z) \, dz \end{gathered}}$$

ex

$$\begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_B \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_0 \\ \mu_1 \\ \mu_2 \\ \vdots \\ \mu_B \end{bmatrix}, \begin{bmatrix} \quad & \quad \\ \quad & \quad \end{bmatrix} \right)$$

WE ONLY CARE ABOUT THE TRAIN AND TEST SET, SO WE MARGINALIZE THE DISTRIBUTION AND OBTAIN A NEW ONE. WE CAN DEFINE THE NON-PARAMETRIC PREDICTION (LINEAR REGRESSION) AS FOLLOWS

$$\boxed{y^{(i)} = f(x^{(i)}) + \varepsilon^{(i)}}, \quad i = 1, \dots, n$$

DISTR. FUNCTION, NOT DIRECTLY THE PARAMETER

WITH PARAMETERS BEING DEFINED AS FOLLOWS

TRAIN SET

$$\underline{X} = \begin{bmatrix} - & x^{(1)} & - \\ & \vdots & \\ - & x^{(n)} & - \end{bmatrix} \in \mathbb{R}^{n \times d} \quad \vec{f} = \begin{bmatrix} f(x^{(1)}) \\ \vdots \\ f(x^{(n)}) \end{bmatrix} \quad \vec{\varepsilon} = \begin{bmatrix} \varepsilon^{(1)} \\ \vdots \\ \varepsilon^{(n)} \end{bmatrix} \quad \underline{\vec{y}} = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(n)} \end{bmatrix} \in \mathbb{R}^n$$

X, y KNOWN

TEST SET

$$\underline{X_*} = \begin{bmatrix} — x_*^{(1)} — \\ \vdots \\ — x_*^{(n)} — \end{bmatrix} \in \mathbb{R}^{n \times d} \quad \vec{f} = \begin{bmatrix} f(x_*^{(1)}) \\ \vdots \\ f(x_*^{(n)}) \end{bmatrix} \quad \vec{\varepsilon}_* = \begin{bmatrix} \varepsilon_*^{(1)} \\ \vdots \\ \varepsilon_*^{(n)} \end{bmatrix} \quad \underline{\vec{y}_*} = \begin{bmatrix} y_*^{(1)} \\ \vdots \\ y_*^{(n)} \end{bmatrix} \in \mathbb{R}$$

X KNOWN
y TO PREDICT

WE CAN OBSERVE THAT

$$f(x_*) = \mathbb{E}[y|x] \quad \text{(MEAN 0)}$$

DETERMINISTIC

LET'S ASSUME THAT $f \sim N(0, K(\cdot,\cdot))$

WE NEED THE FULL DISTRIBUTION $y|x$. BY MARGINALIZING A GAUSSIAN PROCESS WE OBTAIN A M-NGD, AS FOLLOWS

$$\begin{bmatrix} \vec{f} \\ \vec{f}_* \end{bmatrix} \Big| x, x_* \sim N\left(\vec{0}, \begin{bmatrix} K(x,x) & K(x,x_*) \\ K(x_*,x) & K(x_*,x_*) \end{bmatrix}\right) \quad (*)$$

NOW WE NEED THE ESTIMATES OF THE NOISE

$$(*) + \varepsilon = \begin{bmatrix} \vec{y} \\ \vec{y}_* \end{bmatrix}$$

NOW WE CAN FINALLY COME TO THE JOINT DISTRIBUTION $\vec{y}, \vec{y}_*$ AND FINALLY OBTAIN A POSTERIOR PREDICTIVE DISTRIBUTION (NON PARAMETRIC)

$$\underbrace{\begin{bmatrix} \vec{y} \\ \vec{y}_* \end{bmatrix}}^{\text{JOINT}} \Big| x, x_* = \overbrace{\begin{bmatrix} \vec{f} \\ \vec{f}_* \end{bmatrix}}^{\Sigma_f = K_{ij}} + \overbrace{\begin{bmatrix} \varepsilon \\ \varepsilon_* \end{bmatrix}}^{\Sigma_\varepsilon = \sigma^2 I} \overset{\text{SUMMATION PROPRIETY}}{=\!=} N\left(\vec{0} + \vec{0}, \begin{bmatrix} K(x,x) + \sigma^2 I & K(x,x_*) \\ K(x_*,x) & K(x_*,x_*) + \sigma^2 I \end{bmatrix}\right)$$

LET'S NOW APPLY THE CONDITIONING PROPRIETY, AND WE OBTAIN THE POSTERIOR PREDICTIVE DISTRIBUTION

$$\boxed{\vec{y}_* \mid \vec{y}, x, x_* \sim N(\mu_*, \Sigma_*)} \xrightarrow{\text{PARAMS}} \begin{cases} \mu_* = K(x, x_*)\left(K(x,x)+\sigma^2 I\right)^{-1} \vec{y} \\ \Sigma^* = K(x_*, x) + \sigma^2 I - K(x_*, x)\left(K(x,x)+\sigma^2 I\right)^{-1} K(x, x_*) \end{cases}$$

$$= X_*^T X \left(X^T X + \sigma^2 I\right)^{-1} \vec{y}$$

LET'S RECALL SOME CONCEPTS ABOUT POSTERIOR DISTRIBUTIONS

→ IF OVER $\vec{y}$: PREDICTIVE

→ IF OVER OTHER PARAMS SUCH AS θ: NON-PREDICTIVE

REMEMBER THAT WE HAVEN'T MADE ANY ASSUMPTION ON $\vec{f}$ AT ALL, THAT COULD BE "VERY NON-LINEAR"

GPs ARE PRETTY EXPENSIVE BUT VERY GOOD IN PREDICTING $(y \in \mathbb{R})$

WE CAN HAVE <u>NO</u> PARAMETRIC COSTRAINT, BUT STILL NEED TO CHOOSE A KERNEL FUNCTION (EVERYTHING DERIVES FROM THAT CONSEQUENTIALLY)