

# Processus Gaussien

Cours 8  
ML Master DAC

Nicolas Baskiotis  
nicolas.baskiotis@sorbonne-universite.fr

équipe MLIA, Institut des Systèmes Intelligents et de Robotique (ISIR)  
Sorbonne Université

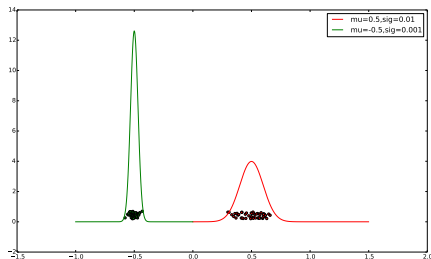
S2 (2021-2022)

# Plan

- 1 **Digression : Gaussiennes multivariées**
- 2 Retour sur la régression
- 3 La magie de la gaussienne
- 4 Processus Gaussien pour la régression

# Rappel : distribution gaussienne

- En 1d :  $p(x) = \mathcal{N}(x|\mu, \sigma) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{(-\frac{1}{2\sigma^2}(x-\mu)^2)}$



Remarque : à quoi sert la constante :  $\frac{1}{(2\pi\sigma^2)^{1/2}}$  ?

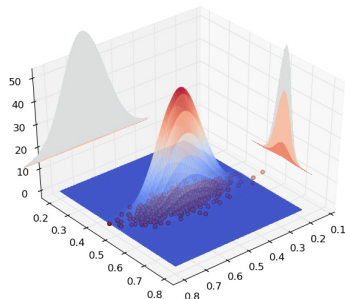
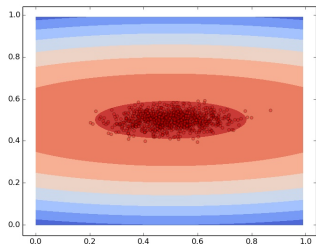
- Multivariée en  $d$  dimensions:

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x - \mu)' \Sigma^{-1} (x - \mu)\right)$$

- $\mu = (\mu_1, \mu_2, \dots, \mu_d)$ , mais  $\Sigma$  ?

# Gaussienne 2D : cas simple

- En 2d : on suppose que  $x_1 \sim \mathcal{N}(\mu_1, \sigma_1)$  et  $x_2 \sim \mathcal{N}(\mu_2, \sigma_2)$
- hypothèse Naive Bayes,  $x_1$  indépendant de  $x_2$
- $p(x) = p(x|\mathcal{N}_1)p(x|\mathcal{N}_2) = \frac{1}{(2\pi\sigma_1^2)^{1/2}} e^{-\frac{1}{2\sigma_1^2}(x_1-\mu_1)^2} \frac{1}{(2\pi\sigma_2^2)^{1/2}} e^{-\frac{1}{2\sigma_2^2}(x_2-\mu_2)^2}$
- $p(x) = \frac{1}{(2\pi)^{2/2}(\sigma_1^2\sigma_2^2)^{1/2}} e^{-\frac{1}{2}\left(\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2}\right)} = \frac{1}{2\pi\Sigma^{1/2}} e^{-\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)}$   
avec  $\Sigma = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$



# Gaussienne 2D : cas générique

## Transformation affine

- Supposons  $x_1, x_2 \sim \mathcal{N}(0, 1)$  et  $X = (x_1, x_2)$ ;
- Soit  $T$  une transformation affine inversible  $T = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix}$
- Soit  $Y = TX + \mu$ ,  $y_1 = t_{11}x_1 + t_{12}x_2 + \mu_1$ ,  $y_2 = t_{21}x_1 + t_{22}x_2 + \mu_2$
- alors  $\mathbb{E}(Y) = \begin{pmatrix} \mathbb{E}(t_{11}x_1 + t_{12}x_2 + \mu_1) \\ \mathbb{E}(t_{21}x_1 + t_{22}x_2 + \mu_2) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$
- Variance d'un vecteur aléatoire ?

# Covariance

## Covariance dans le cas général

- Covariance de deux variables aléatoires :

$$\text{Cov}(x_1, x_2) = \mathbb{E}((x_1 - \mathbb{E}(x_1))(x_2 - \mathbb{E}(x_2))) = \mathbb{E}(x_1 x_2) - \mathbb{E}(x_1)\mathbb{E}(x_2)$$

- Matrice de covariance d'un vecteur aléatoire  $X$ ,  $\text{Cov}(X)$  :

$$\begin{pmatrix} \text{Cov}(x_1, x_1) & \cdots & \text{Cov}(x_1, x_n) \\ \vdots & \ddots & \vdots \\ \text{Cov}(x_n, x_1) & \cdots & \text{Cov}(x_n, x_n) \end{pmatrix} = \mathbb{E}((X - \mu)(X - \mu)') = \mathbb{E}(XX') - \mu\mu'$$

## Gaussienne 2D : cas générique

On suppose toujours :  $x_1, x_2 \sim \mathcal{N}(0, 1)$ ,  $X = (x_1, x_2)$  et  $T = \begin{pmatrix} t_{11} & t_{12} \\ t_{21} & t_{22} \end{pmatrix}$  et

$$Y = TX + \mu$$

**Covariance de  $Y$  :**  $Cov(Y) = Cov(TX + \mu) = TT'$

$$\begin{aligned} \bullet \text{ Cov}(Y) &= \begin{pmatrix} \mathbb{E}((t_{11}x_1 + t_{12}x_2)^2) & \mathbb{E}((t_{11}x_1 + t_{12}x_2)(t_{21}x_1 + t_{22}x_2)) \\ \mathbb{E}((t_{11}x_1 + t_{12}x_2)(t_{21}x_1 + t_{22}x_2)) & \mathbb{E}((t_{21}x_1 + t_{22}x_2)^2) \end{pmatrix} \\ &= \begin{pmatrix} t_{11}^2 + t_{12}^2 & t_{11}t_{21} + t_{12}t_{22} \\ t_{11}t_{21} + t_{12}t_{22} & t_{21}^2 + t_{22}^2 \end{pmatrix} \end{aligned}$$

On note  $\Sigma = Cov(Y)$

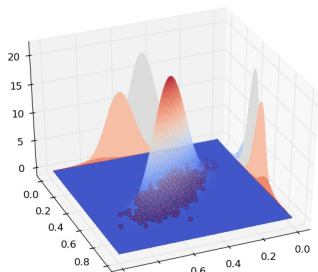
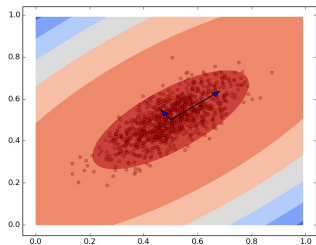
## Changement de variable

- $\bullet p(x) = \frac{1}{2\pi \Sigma_{\mathcal{N}(0,1)}^{1/2}} e^{-\frac{1}{2}(x - \mu_{\mathcal{N}(0,1)})' \Sigma_{\mathcal{N}(0,1)}^{-1} (x - \mu_{\mathcal{N}(0,1)})}$ , avec  $\mu_{\mathcal{N}(0,1)} = 0$ ,  $\Sigma_{\mathcal{N}(0,1)} = I$
- $\bullet$  Si  $Y = TX + \mu$ , alors  $p(Y) = \frac{1}{|\det(T)|} p(T^{-1}(Y - \mu))$
- $\bullet p(Y) = \frac{1}{2\pi |\Sigma|^{-1/2}} e^{-\frac{1}{2}((T^{-1}(Y - \mu))' IT^{-1}(Y - \mu))} = \frac{1}{|\Sigma|^{-1/2} 2\pi} e^{-\frac{1}{2}(Y - \mu)' T'^{-1} T^{-1} (Y - \mu)}$
- $\bullet p(Y) = \frac{1}{2\pi |\Sigma|^{-1/2}} e^{\frac{1}{2}(Y - \mu)' \Sigma^{-1} (Y - \mu)}$

# Gaussienne 2D : interprétation géométrique

## Transformation affine inversible

- $T$  peut être décomposé en  $T = UD$ ,  $D$  diagonale (valeurs propres) et  $U$  orthogonale (vecteurs propres, matrice de rotation et réflexion)
  - $\Sigma = UD(UD)' = UDD'U' = UD^2U'$
  - $Det(\Sigma) = Det(UD^2U') = Det(D^2) = \sum_i \sigma_i^2$ ,  $\sigma_i$  valeurs propres de  $T$
- ⇒ Loi normale multivariée :  $D$  représente la variance sur chaque composante normale indépendante des autres,  
 $U$  représente la rotation/reflexion par rapport aux axes.





# Plan

- 1 Digression : Gaussiennes multivariées
- 2 Retour sur la régression**
- 3 La magie de la gaussienne
- 4 Processus Gaussien pour la régression

# Régression et noyaux

## Formulation

Pour un jeu de données  $\mathcal{D} = \{\mathbf{x}^i, y^i\}_{i=1}^N$   $\mathbf{x}^i \in \mathbb{R}^d, y^i \in \mathbb{R}$  i.i.d.

- On se donne un noyau  $K(\mathbf{x}^1, \mathbf{x}^2) = \langle \phi(\mathbf{x}^1), \phi(\mathbf{x}^2) \rangle$ , avec  $\phi : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$
- Régression pénalisée :  $J(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (\mathbf{w}^t \phi(\mathbf{x}^i) - y^i)^2 + \frac{\lambda}{2} \mathbf{w}^t \mathbf{w}$
- On note  $\Phi \in \mathbb{R}^{N \times d'}$  la matrice des  $\phi(\mathbf{x}^i)$ ,

$$J(\mathbf{w}) = \frac{1}{2} (\mathbf{w}^t \Phi^t - \mathbf{y}^t) (\mathbf{w}^t \Phi^t - \mathbf{y}^t)^t + \frac{\lambda}{2} \mathbf{w}^t \mathbf{w} = \frac{1}{2} \mathbf{w}^t \Phi^t \Phi \mathbf{w} - \mathbf{w}^t \Phi^t \mathbf{y} + \frac{1}{2} \mathbf{y}^t \mathbf{y} + \frac{\lambda}{2} \mathbf{w}^t \mathbf{w}$$

# Régression et noyaux

## Annulation du gradient

$$J(\mathbf{w}) = \frac{1}{2}(\mathbf{w}^t \Phi^t - \mathbf{y}^t)(\mathbf{w}^t \Phi^t - \mathbf{y}^t)^t + \frac{\lambda}{2} \mathbf{w}^t \mathbf{w} = \frac{1}{2} \mathbf{w}^t \Phi^t \Phi \mathbf{w} - \mathbf{w}^t \Phi^t \mathbf{y} + \frac{1}{2} \mathbf{y}^t \mathbf{y} + \frac{\lambda}{2} \mathbf{w}^t \mathbf{w}$$

\*

- Annulation du gradient par rapport à  $\mathbf{w}$  :

$$\nabla_{\mathbf{w}} J(\mathbf{w}) = \Phi^t \Phi \mathbf{w} - \Phi^t \mathbf{y} + \lambda \mathbf{w} = 0$$

$$\mathbf{w} = -\frac{1}{\lambda} \Phi^t (\Phi \mathbf{w} - \mathbf{y}) = \Phi^t \mathbf{a}$$

- En ré-écrivant :

$$J(\mathbf{a}) = \frac{1}{2} \mathbf{a}^t \Phi \Phi^t \Phi^t \Phi \mathbf{a} - \mathbf{a}^t \Phi \Phi^t \mathbf{y} + \frac{1}{2} \mathbf{y}^t \mathbf{y} + \frac{\lambda}{2} \mathbf{a}^t \Phi \Phi^t \mathbf{a}$$

# Régression et noyaux

$$J(\mathbf{a}) = \frac{1}{2} \mathbf{a}^t \Phi \Phi^t \Phi^t \Phi \mathbf{a} - \mathbf{a}^t \Phi \Phi^t \mathbf{y} + \frac{1}{2} \mathbf{y}^t \mathbf{y} + \frac{\lambda}{2} \mathbf{a}^t \Phi \Phi^t \mathbf{a}$$

## Résolution

- On note  $K = \Phi^T \Phi$ , avec  $K_{i,j} = \langle \phi(\mathbf{x}^i), \phi(\mathbf{x}^j) \rangle = k(\mathbf{x}^i, \mathbf{x}^j)$ , la matrice de Gram du noyau (symétrique), on a

$$J(\mathbf{a}) = \frac{1}{2} \mathbf{a}^t K^t K \mathbf{a} - \mathbf{a}^t K^t \mathbf{y} + \frac{1}{2} \mathbf{y}^t \mathbf{y} + \frac{\lambda}{2} \mathbf{a}^t K^t \mathbf{a}$$

- En prenant le gradient par rapport à  $\mathbf{a}$ , on trouve

$$\mathbf{a} = (K + \lambda I_N)^{-1} \mathbf{y}$$

- Pour la prédiction en  $\mathbf{x}$  :  $\mathbf{w}^t \phi(\mathbf{x}) = \mathbf{a}^t \Phi \phi(\mathbf{x}) = ((K + \lambda I_N)^{-1} \mathbf{y})^t \mathbf{k}(\mathbf{x})^t$ , avec  $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x}^1, \mathbf{x}), \dots, k(\mathbf{x}^N, \mathbf{x}))$

# Hypothèse du bruit gaussien

## Formalisation

- Un jeu de données  $\mathcal{D} = \{\mathbf{x}^i, y^i\}_{i=1}^N$   $\mathbf{x}^i \in \mathbb{R}^d, y^i \in \mathbb{R}$  i.i.d.
  - On suppose que  $y^i = f(\mathbf{x}^i) + \epsilon^i$
  - Régression linéaire :  $f$  de la forme  $\mathbf{w}^t \cdot \mathbf{x}$  (on oublie le biais pour simplifier)
  - Le petit "plus" par rapport à la régression linéaire "simple" : on suppose  $\epsilon^i \sim \mathcal{N}(0, \sigma^2)$ , indépendant de  $\mathbf{x}^i$
- ⇒ la distribution de  $y$  conditionnée au modèle et à l'entrée  $\mathbf{x}$  est gaussienne :

$$p(y^i | \mathbf{x}^i; \mathbf{w}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^i - \mathbf{w}^t \cdot \mathbf{x}^i)^2}{2\sigma^2}}$$

# Hypothèse du bruit gaussien

## Résolution par maximum de vraisemblance

$$L(\mathbf{w}, \sigma^2) = \prod_{i=1}^N p(y^i | \mathbf{x}^i; \mathbf{w}, \sigma^2) = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y^i - \mathbf{w}^t \cdot \mathbf{x}^i)^2}{2\sigma^2}}$$

$$\log L(\mathbf{w}, \sigma^2) = -\frac{N}{2} \log 2\pi - N \log(\sigma) - \frac{1}{2\sigma^2} \sum_{i=1}^N (y^i - \mathbf{w}^t \cdot \mathbf{x}^i)^2$$

→ gradient par rapport à  $\mathbf{w}$  :  $\mathbf{w} = (X^T X)^{-1} X^t \mathbf{y}$

→ gradient par rapport à  $\sigma$  :  $\sigma^2 = \frac{1}{N} \sum_{i=1}^N (y^i - \mathbf{w}^t \cdot \mathbf{x}^i)^2$

## Quelle est l'avantage alors de rajouter l'hypothèse de bruit gaussien ?

⇒ On connaît pour une entrée  $\mathbf{x}$  la distribution de l'estimée  $\hat{y}$  qui suit une loi gaussienne ...

# Plan

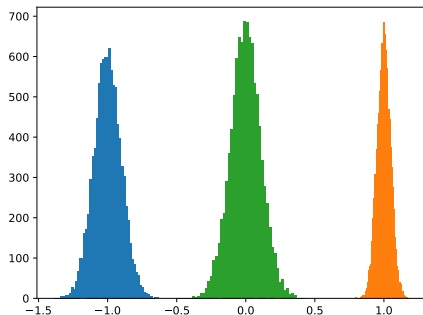
- 1 Digression : Gaussiennes multivariées
- 2 Retour sur la régression
- 3 La magie de la gaussienne**
- 4 Processus Gaussien pour la régression

# Une gaussienne et tout est gaussien !

## Somme de gaussiennes :

Soit  $\mathbf{y}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$  et  $\mathbf{y}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_2)$ ,  
alors la variable aléatoire  $\mathbf{y} = \mathbf{y}_1 + \mathbf{y}_2$  suit une loi normale :

$$\mathcal{N}(\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2, \Sigma_1 + \Sigma_2)$$





# Une gaussienne et tout est gaussien !

## Marginalisation

Soit  $\mathbf{y} = (y_1, \dots, y_d) \in \mathbb{R}^d$ ,  $\mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ ,  $p(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})' \Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})}$

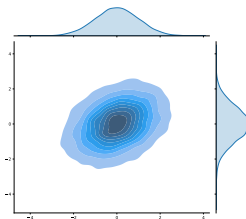
On considère une partition en 2 groupes :

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

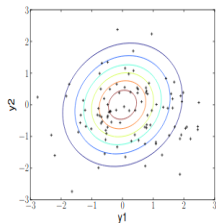
avec  $\Sigma_{11}$  et  $\Sigma_{22}$  carrés symétriques (et donc  $\Sigma_{12}' = \Sigma_{21}$ ).

Alors la marginalisation est gaussienne :

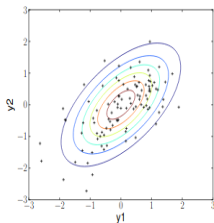
$$p(\mathbf{y}_1) = \int_{\mathbf{y}_2} p(\mathbf{y}_1, \mathbf{y}_2; \boldsymbol{\mu}, \Sigma) d\mathbf{y}_2 = \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11})$$



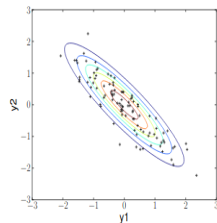
# Effet de la covariance



$$\Sigma = \begin{bmatrix} 1 & 0.14 \\ 0.14 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & -0.9 \\ -0.9 & 1 \end{bmatrix}$$

# Une gaussienne et tout est gaussien !

## Conditionnement

$$\text{Soit } \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{bmatrix}, \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

La conditionnée est gaussienne :

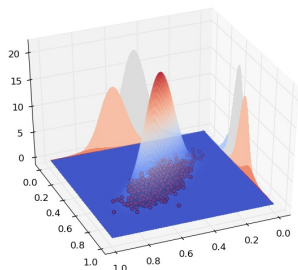
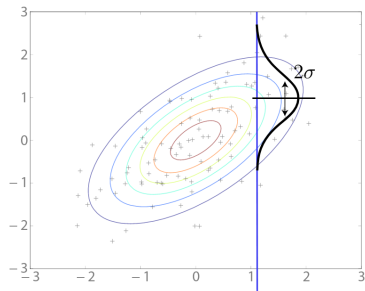
$$p(\mathbf{y}_2 | \mathbf{y}_1 = \mathbf{a}; \boldsymbol{\mu}, \Sigma) = \frac{p(\mathbf{y}; \boldsymbol{\mu}, \Sigma)}{\int_{\mathbf{y}_2} p(\mathbf{y}; \boldsymbol{\mu}, \Sigma) d\mathbf{y}_2}$$

et suit la loi

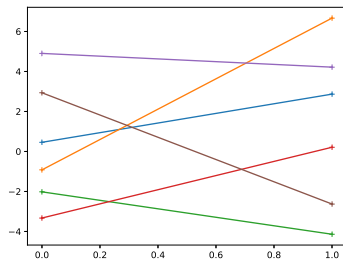
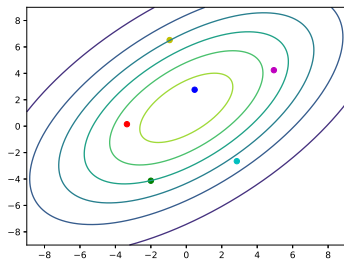
$$\mathcal{N}(\boldsymbol{\mu}_2 + \Sigma_{21} \Sigma_{11}^{-1} (\mathbf{a} - \boldsymbol{\mu}_1), \Sigma_{22} - \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12})$$

# Corrélation entre coordonnées

$$p(\mathbf{y}_2 | \mathbf{y}_1 = \mathbf{a}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \sim \mathcal{N}(\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{a} - \boldsymbol{\mu}_1), \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})$$

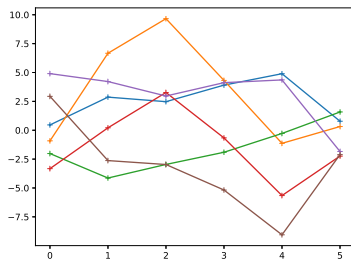
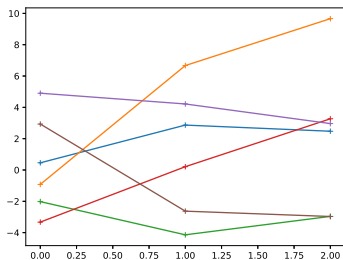


# Une autre manière de visualiser une gaussienne



Pour chaque point 2D tiré de cette gaussienne, la première coordonnée est placée en 0, sa deuxième en 1.

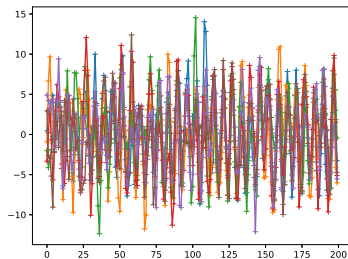
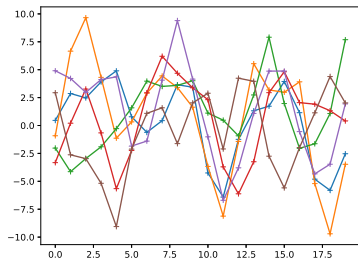
# En 3d et plus ...



Soit une gaussienne en  $N$  dimensions, la  $i$ -ème coordonnée est placée en  $x = i$ .

(rappel : la relation entre la  $i$ -ème et  $j$ -ème dimension est définie par la covariance  $\Sigma_{ij}$ )

## En 3d et plus ...



Soit une gaussienne en  $N$  dimensions, la  $i$ -ème coordonnée est placée en  $x = i$ .

(rappel : la relation entre la  $i$ -ème et  $j$ -ème dimension est définie par la covariance  $\Sigma_{ij}$ )

# Plan

- 1 Digression : Gaussiennes multivariées
- 2 Retour sur la régression
- 3 La magie de la gaussienne
- 4 Processus Gaussien pour la régression**



# Régression linéaire bayésienne

Rappel :  $f(\mathbf{x}) = \mathbf{w}^t \phi(\mathbf{x})$  (déterministe) et  $y = f(\mathbf{x}) + \epsilon$  (avec  $\epsilon$  bruit gaussien) et des données  $D = \{\mathbf{x}^i, y^i\}_{i=1}^N$

Processus habituel : de  $D$  on estime  $\mathbf{w}$ , puis on fait les prédictions.

## Où sont les gaussiennes ?

- $p(y|\mathbf{x}; \mathbf{w})$  : gaussien
- La vraisemblance :  $p(D|\mathbf{w}) = \prod_{i=1}^N p(y^i|\mathbf{x}^i; \mathbf{w}) \Rightarrow$  gaussien
- Le prior :  $p(\mathbf{w})$  est gaussien (dans le cadre de la ridge régression)
- Le posterior :  $p(\mathbf{w}|D) = \frac{p(\mathbf{w})p(D|\mathbf{w})}{p(D)} \Rightarrow$  gaussien

# Régression linéaire bayésienne

Rappel :  $f(\mathbf{x}) = \mathbf{w}^t \phi(\mathbf{x})$  (déterministe) et  $y = f(\mathbf{x}) + \epsilon$  (avec  $\epsilon$  bruit gaussien) et des données  $D = \{\mathbf{x}^i, y^i\}_{i=1}^N$

Processus habituel : de  $D$  on estime  $\mathbf{w}$ , puis on fait les prédictions.

## Où sont les gaussiennes ?

- $p(y|\mathbf{x}; \mathbf{w})$  : gaussien
- La vraisemblance :  $p(D|\mathbf{w}) = \prod_{i=1}^N p(y^i|\mathbf{x}^i; \mathbf{w}) \Rightarrow$  gaussien
- Le prior :  $p(\mathbf{w})$  est gaussien (dans le cadre de la ridge régression)
- Le posterior :  $p(\mathbf{w}|D) = \frac{p(\mathbf{w})p(D|\mathbf{w})}{p(D)} \Rightarrow$  gaussien

## Mais si on se passait de l'estimation de $\mathbf{w}$ :

- Ce qu'on veut :  $p(y|\mathbf{x}, D)$
  - Mais  $p(y|\mathbf{x}, D) = \int_{\mathbf{w}} p(y|\mathbf{x}; \mathbf{w})p(\mathbf{w}|D)d\mathbf{w}$
  - Or tous les termes sont gaussiens
- $\Rightarrow p(y|\mathbf{x}, D)$  est gaussien !
- Donc  $p(y|\mathbf{x}, D) \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ , il suffit d'estimer  $\boldsymbol{\mu}$  et  $\Sigma$  pour prédire  $y$ .

# En résumé

Soit  $D = \{\mathbf{x}^i, y^i\}_{i=1}^N$  nos données et  $\mathbf{x}_t^1 \dots \mathbf{x}_t^T$  les points que l'on veut inférer. On a établi que (en simplifiant en fixant la moyenne à 0) :

$$p \left( \begin{pmatrix} y^1 \\ \vdots \\ y^N \\ y_t^1 \\ \vdots \\ y_t^T \end{pmatrix} \mid \mathbf{x}^1, \mathbf{x}^2, \dots, \mathbf{x}^N, \mathbf{x}_t^1, \dots, \mathbf{x}_t^T \right) \sim \mathcal{N}(0, \Sigma), \quad \Sigma = \begin{bmatrix} K & K_{\star} \\ K_{\star}^t & K_{\star\star} \end{bmatrix}$$

Alors

$$p(y_t^1 \dots y_t^T \mid \mathbf{y}, \mathbf{x}^1, \dots, \mathbf{x}_t^T) \sim \mathcal{N}(K_{\star}^t K^{-1} \mathbf{y}, K_{\star\star} - K_{\star}^t K^{-1} K_{\star})$$

Mais comment obtenir:

- $K$  les covariances entre les points d'entraînement ?
- $K_{\star}$  les covariances entre entraînement et test ?
- $K_{\star\star}$  les covariances entre test ?

# Matrice de covariance = Kernel !

## Ce que l'on veut pour la matrice de covariance :

- qu'elle soit symétrique ! ( $i$  influence  $j$  comme  $j$  influence  $i$ )
- deux points "similaires" doivent avoir une corrélation forte :  $Cov(\mathbf{x}^i, \mathbf{x}^j)$  grand
- deux points "dissimilaires" doivent avoir une corrélation faible :  $Cov(\mathbf{x}^i, \mathbf{x}^j)$  petit
- qu'elle soit semi-défini positive
- $Cov(\mathbf{x}^i, \mathbf{x}^i)$  doit dénoté la variance en ce point

⇒ Très similaire à la notion de noyaux en SVM ! Autant utiliser une fonction noyau pour encoder la covariance ...

## Covariances typiques :

- Squared Exponential :  $K(\mathbf{x}^1, \mathbf{x}^2) = \sigma^2 e^{-\frac{1}{2} \left( \frac{\|\mathbf{x}^1 - \mathbf{x}^2\|^2}{\lambda} \right)}$
- Linéaire :  $K(\mathbf{x}^1, \mathbf{x}^2) = \lambda + \langle \mathbf{x}^1, \mathbf{x}^2 \rangle$
- Periodic :  $K(\mathbf{x}^1, \mathbf{x}^2) = \sigma^2 e^{-\frac{2\sin^2(\frac{\|\mathbf{x}^1 - \mathbf{x}^2\|}{2})}{\lambda^2}}$

# Et si on introduit du bruit ?

## Bruit additif gaussien

- On observe  $y^i = f(\mathbf{x}^i) + \epsilon_i$ , avec  $\epsilon_i$  indépendant et suivant  $\mathcal{N}(0, \sigma^2)$
- La covariance est changée en  $\hat{\Sigma}$ :

$$\hat{\Sigma}_{ij} = \mathbb{E}[(f(\mathbf{x}^i) + \epsilon_i)(f(\mathbf{x}^j) + \epsilon_j)] = \mathbb{E}[f(\mathbf{x}^i)f(\mathbf{x}^j)] + \mathbb{E}[f(\mathbf{x}^j)]\mathbb{E}[\epsilon_i] + \mathbb{E}[f(\mathbf{x}^i)]\mathbb{E}[\epsilon_j] + \mathbb{E}[\epsilon_i\epsilon_j]$$

- Pour  $i \neq j$ ,  $\mathbb{E}[\epsilon_i] = 0$ ,  $\mathbb{E}[\epsilon_i\epsilon_j] = \mathbb{E}[\epsilon_i]\mathbb{E}[\epsilon_j] = 0$

$$\hat{\Sigma}_{ij} = \mathbb{E}[f(\mathbf{x}^i)f(\mathbf{x}^j)] = \Sigma_{ij}$$

- Pour  $i = j$ ,  $\mathbb{E}[\epsilon_i] = 0$ ,  $\mathbb{E}[\epsilon_i^2] = \sigma^2$

$$\hat{\Sigma}_{ii} = \mathbb{E}[f(\mathbf{x}^i)f(\mathbf{x}^i)] + \mathbb{E}[\epsilon_i^2] = \Sigma_{ii} + \sigma^2$$

- Donc  $\hat{\Sigma} = \Sigma + \sigma^2\mathbf{I}$

## Formule de la régression GP dans le cas général

$$p(y_t^1 \dots y_t^T | \mathbf{y}, \mathbf{x}^1, \dots, \mathbf{x}_t^T) \sim \mathcal{N}(K_*^t (K + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, K_{**} - K_*^t (K + \sigma^2 \mathbf{I})^{-1} K_*)$$

⇒ Régression à noyaux !

- Mais avec l'information sur l'incertitude liée à la prédiction !

# Définition formelle des Processus Gaussien (GP)

## Définition

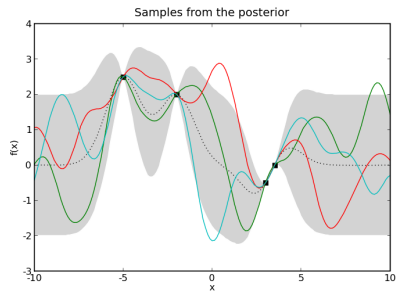
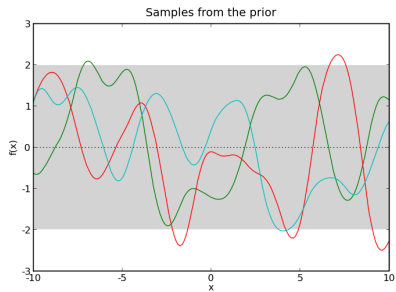
- La fonction  $f$  dont on cherche la prédiction est vue comme une collection de points  $f$  (les mesures en différents points) potentiellement infinie.
- Un processus gaussien est une collection de variables aléatoires (potentiellement infinie) tels que la distribution jointe de tout sous-ensemble de ces variables est une gaussienne multivariée :

$$f \sim GP(\mu, k)$$

avec  $\mu(\mathbf{x})$  et  $k(\mathbf{x}^1, \mathbf{x}^2)$  sont les fonctions de moyenne et de covariance.

- On cherche à estimer la distribution  $P(f_t | \mathbf{x}_t, D)$  en utilisant un prior GP :  $P(f | \mathbf{x}) \sim \mathcal{N}(\mu, \Sigma)$  et en le conditionnant par les données d'entraînement  $D$  afin de modéliser la distribution jointe  $f$  des points d'entraînement et  $f_t$  les points de test.

# Examples



# Conclusion

## Les processus gaussiens

- sont relativement puissants sous certaines conditions
- sont adaptables à beaucoup de tâches (classification, non supervisé, active learning, ...)
- donnent une mesure d'incertitude liée à la prédiction
- Mais temps de calcul en  $O(N^3)$  avec inversion de matrice !
- Les hyper-paramètres sont cachés dans le noyau ...

## Références:

Cours ETHZ

Cours Cornell U., très bonne intro vidéo

Très bon livre de Rasmussen et Williams