

ガウス過程回帰

Gaussian Process Regression

GPR

明治大学 理工学部 応用化学科
データ化学工学研究室 金子 弘昌

ガウス過程による回帰 (GPR) とは？

- ✓ 線形の回帰分析手法
- ✓ カーネルトリックにより非線形の回帰モデルに
- ✓ 目的変数の推定値だけでなく、その分散も計算できる
- ✓ クロスバリデーションがいらぬ

GPRを理解するための大まかな流れ

- ✓前提：モデル構築用のサンプルの数を n とし、 $n+1$ 個目のサンプルの目的変数 y の値を推定したいとする
 - n 個のサンプルについては、 y の値と説明変数 X の値があり、 $n+1$ 個目のサンプルについては、 X の値のみがある
- ✓① 線形のモデルを仮定する
 - $y = X b$ (b : 回帰係数)
- ✓② サンプル間の y の関係は、サンプル間の X の関係によって決まることを示す
- ✓③ カーネルトリックにより非線形モデルに拡張する
- ✓④ y にはノイズ (測定誤差) が含まれていることから、そのノイズの大きさを仮定して、再び ② の関係を求める
- ✓⑤ ④から n 個のサンプルの X と、 $n+1$ 個目のサンプルの X との間の関係を求め、さらに n 個の y の値を用いて、 $n+1$ 個目の y の推定値を限定していく

説明に入る前に：GPRがとっつきにくい理由

✓ y と b については、1つの値ではなく分布を考えなければならない

- 具体的には、正規分布 (ガウス分布)
→ “ガウス”過程の名前の由来
- x については、値で OK

… p. 6, 7, 8 で説明

✓ 分布からのサンプリングを理解しなければならない

… p. 13, 14, 15 で説明

- そういう意味では、②が最難関であり、そこを理解して抜けるとそのあとは霧が晴れたように GPR を理解できると思います

① 線形モデルの仮定

$$\checkmark \mathbf{y} = \mathbf{X}\mathbf{b}$$

$$\begin{array}{c} \mathbf{y} \\ \left[\begin{array}{c} y^{(1)} \\ \vdots \\ y^{(i)} \\ \vdots \\ y^{(n)} \end{array} \right] \end{array} = \begin{array}{c} \mathbf{X} \\ \left[\begin{array}{cccc} x_1^{(1)} & \cdots & x_j^{(1)} & \cdots & x_m^{(1)} \\ \vdots & & \vdots & & \vdots \\ x_1^{(i)} & \cdots & x_j^{(i)} & \cdots & x_m^{(i)} \\ \vdots & & \vdots & & \vdots \\ x_1^{(n)} & \cdots & x_j^{(n)} & \cdots & x_m^{(n)} \end{array} \right] \end{array} \begin{array}{c} \mathbf{b} \\ \left[\begin{array}{c} b_1 \\ \vdots \\ b_j \\ \vdots \\ b_m \end{array} \right] \end{array}$$

n : サンプル数

m : 説明変数の数

① 簡単にするため、まずは X を1変数とする

$$\checkmark \mathbf{y} = \mathbf{x}b$$

$$\mathbf{y} \quad \mathbf{x} \quad b$$

$$\begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(i)} \\ \vdots \\ \vdots \\ y^{(n)} \end{bmatrix} = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(i)} \\ \vdots \\ \vdots \\ x^{(n)} \end{bmatrix} b$$

② 回帰係数が正規分布に従うと仮定

- ✓ b の分布を正規分布 (ガウス分布) と仮定する
 - 平均 : 0、分散 : σ_b^2
 - ざっくりいうと、 b は 0.1 かもしれないし、-0.4 かもしれないし、いろいろな可能性がある、ということ

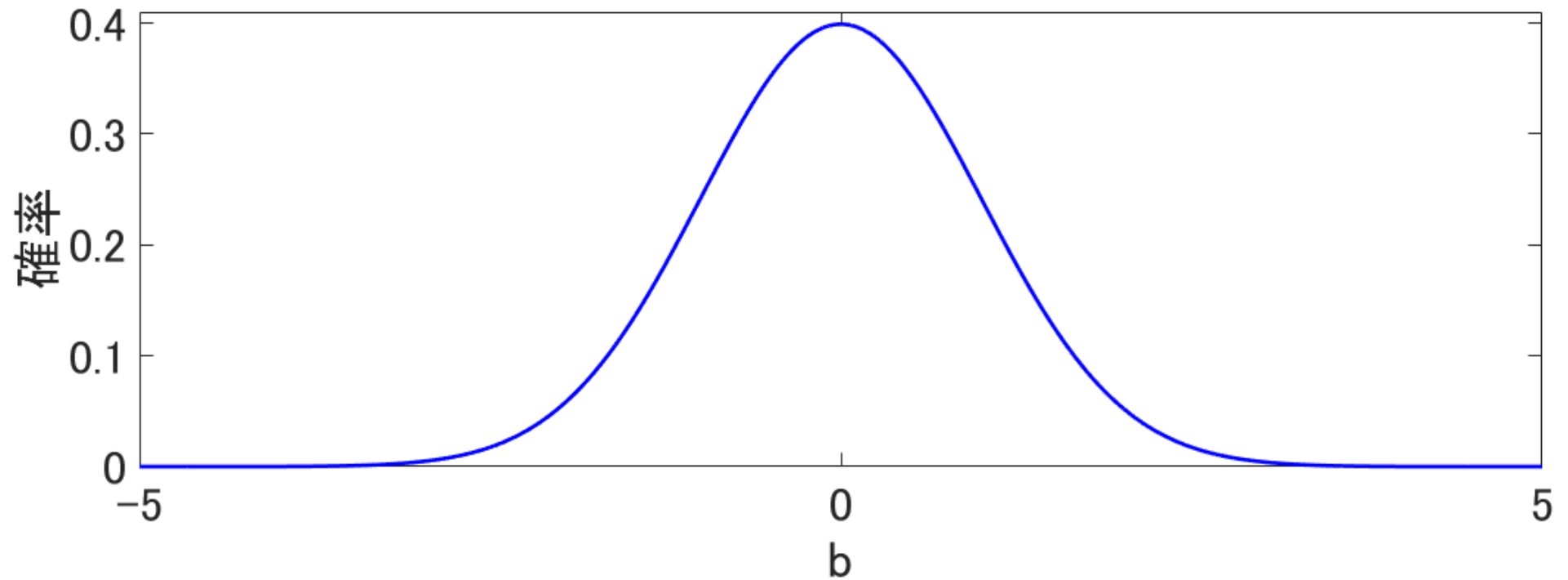
$$p(b) = N(b | 0, \sigma_b^2)$$



b の確率分布 (probability distribution) は正規分布 (Normal distribution) であり、平均 : 0、分散 : σ_b^2 である、という意味

② b の例

$\sqrt{\sigma_b} = 1$ のときの、b の分布



② サンプル間の y の関係を考える

- ✓ 念頭にあること : x の値が似ている (近い) サンプル同士は、
 y の値も似ている (近い) だろう
→ サンプル間における y の値の関係は、
 x の値の関係から計算できるだろう
- ✓ b は 1 つの値ではなく、正規分布として与えられた
→ あるサンプルの y の値 ($y^{(i)}$) も同じように、1 つの値ではなく、
正規分布で与えられる！
 - n 個のサンプルがあるので、 n 個の正規分布
- ✓ n 個の正規分布それぞれの、平均と分散を求めればOK? → No!!
- ✓ 念頭にあった、“サンプル間における y の値の関係”、つまり、
正規分布同士の関係も求める必要がある
→ 共分散

② y の平均ベクトルと分散共分散行列

✓ n 個のサンプルの y における正規分布について、

- $y^{(i)}$ の正規分布の平均を m_i とする
- $y^{(i)}$ の正規分布の分散を σ_{yi}^2 とする
- $y^{(i)}$ の正規分布と $y^{(j)}$ の正規分布との共分散を $\sigma_{yi,j}^2$ とする
 - σ_{yi} は $\sigma_{yi,i}$ と同じ

平均ベクトル \mathbf{m}

$$\mathbf{m} = \begin{bmatrix} m_1 \\ \vdots \\ m_i \\ \vdots \\ m_n \end{bmatrix}$$

分散共分散行列 Σ

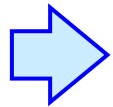
$$\Sigma = \begin{bmatrix} \sigma_{y1,1}^2 & \cdots & \sigma_{y1,j}^2 & \cdots & \sigma_{y1,n}^2 \\ \vdots & & \vdots & & \vdots \\ \sigma_{yi,1}^2 & \cdots & \sigma_{yi,j}^2 & \cdots & \sigma_{yi,n}^2 \\ \vdots & & \vdots & & \vdots \\ \sigma_{yn,1}^2 & \cdots & \sigma_{yn,j}^2 & \cdots & \sigma_{yn,n}^2 \end{bmatrix}$$

② 平均ベクトルと分散共分散行列の計算

✓ $y = \mathbf{x}b$ から、 i 番目のサンプルについては $y^{(i)} = x^{(i)}b$

✓ b の平均は0、分散は σ_b^2

$$m_i = E\left[y^{(i)}\right] = E\left[x^{(i)}b\right] = x^{(i)}E[b] = 0$$



$$\begin{aligned}\sigma_{y^{(i)}, y^{(j)}}^2 &= \text{cov}\left[y^{(i)}, y^{(j)}\right] = \text{cov}\left[x^{(i)}b, x^{(j)}b\right] \\ &= x^{(i)}x^{(j)} \underbrace{\text{cov}[b, b]}_{= b \text{ の分散}} = x^{(i)}x^{(j)}\sigma_b^2\end{aligned}$$

$E[*]$: * の平均

$\text{cov}[* , \cdot]$: * と \cdot との間の共分散

② y の平均ベクトルと分散共分散行列 まとめ ¹¹

$$\mathbf{m} = \begin{bmatrix} m_1 \\ \vdots \\ m_i \\ \vdots \\ m_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{y1,1}^2 & \cdots & \sigma_{y1,j}^2 & \cdots & \sigma_{y1,n}^2 \\ \vdots & & \vdots & & \vdots \\ \sigma_{yi,1}^2 & \cdots & \sigma_{yi,j}^2 & \cdots & \sigma_{yi,n}^2 \\ \vdots & & \vdots & & \vdots \\ \sigma_{yn,1}^2 & \cdots & \sigma_{yn,j}^2 & \cdots & \sigma_{yn,n}^2 \end{bmatrix}$$

$$= \sigma_b^2 \begin{bmatrix} x^{(1)} x^{(1)} & \cdots & x^{(1)} x^{(j)} & \cdots & x^{(1)} x^{(n)} \\ \vdots & & \vdots & & \vdots \\ x^{(i)} x^{(1)} & \cdots & x^{(i)} x^{(j)} & \cdots & x^{(i)} x^{(n)} \\ \vdots & & \vdots & & \vdots \\ x^{(n)} x^{(1)} & \cdots & x^{(n)} x^{(j)} & \cdots & x^{(n)} x^{(n)} \end{bmatrix}$$

② 何を意味するか？

✓ y のサンプル間の分布の関係が、 x のサンプル間の関係で表せた

y の同時分布

✓ x について値が 1 つ与えられると、 y の同時分布が決まる

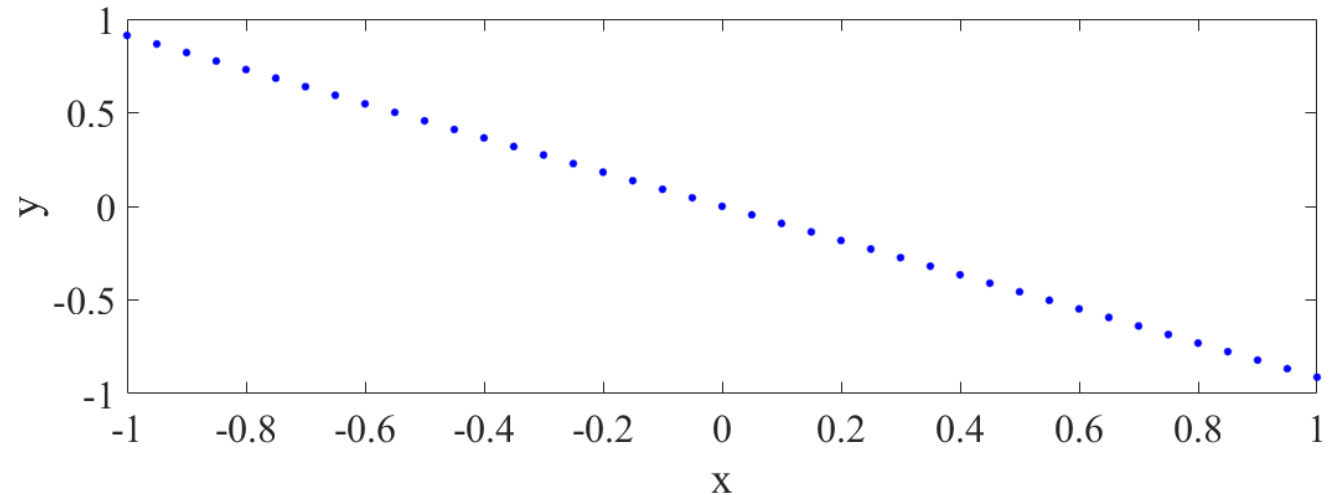
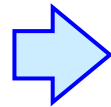
✓ さらに、 b の値が (分布の中から) 1 つに決まると、 y の値が 1 つに決まる

② サンプルを生成してみる

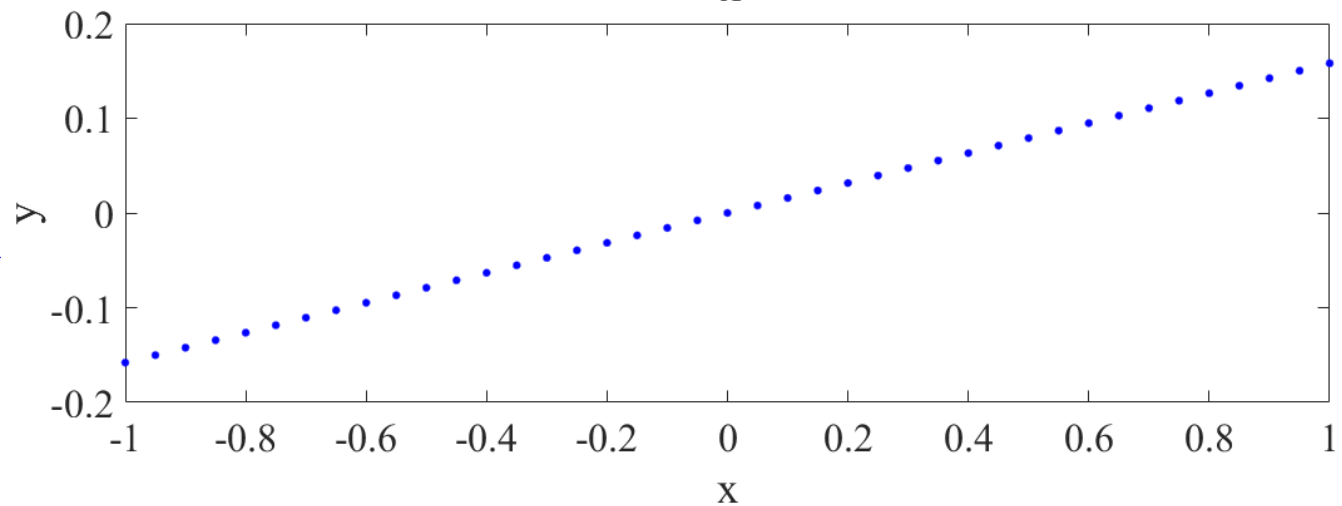
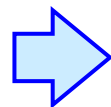
✓ x を、 $-1, -0.95, -0.9, \dots, 0.9, 0.95, 1$ とする

✓ $\sigma_b = 1$ とする

b の値が 1 つに
決まる

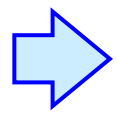


b の値が 1 つに
決まる



② サンプルング

- ✓ 実際は、 b は分布であり、” b の値が 1 つに決まる ” ことに意味はない
- ✓ ただ、 b の値が決まらないと、プロットできない・・・



平均が 0、分散が σ_b^2 の正規分布に従うように、
数多くの b の値を適当に(=ランダムに) 選ぶ
→ サンプルング

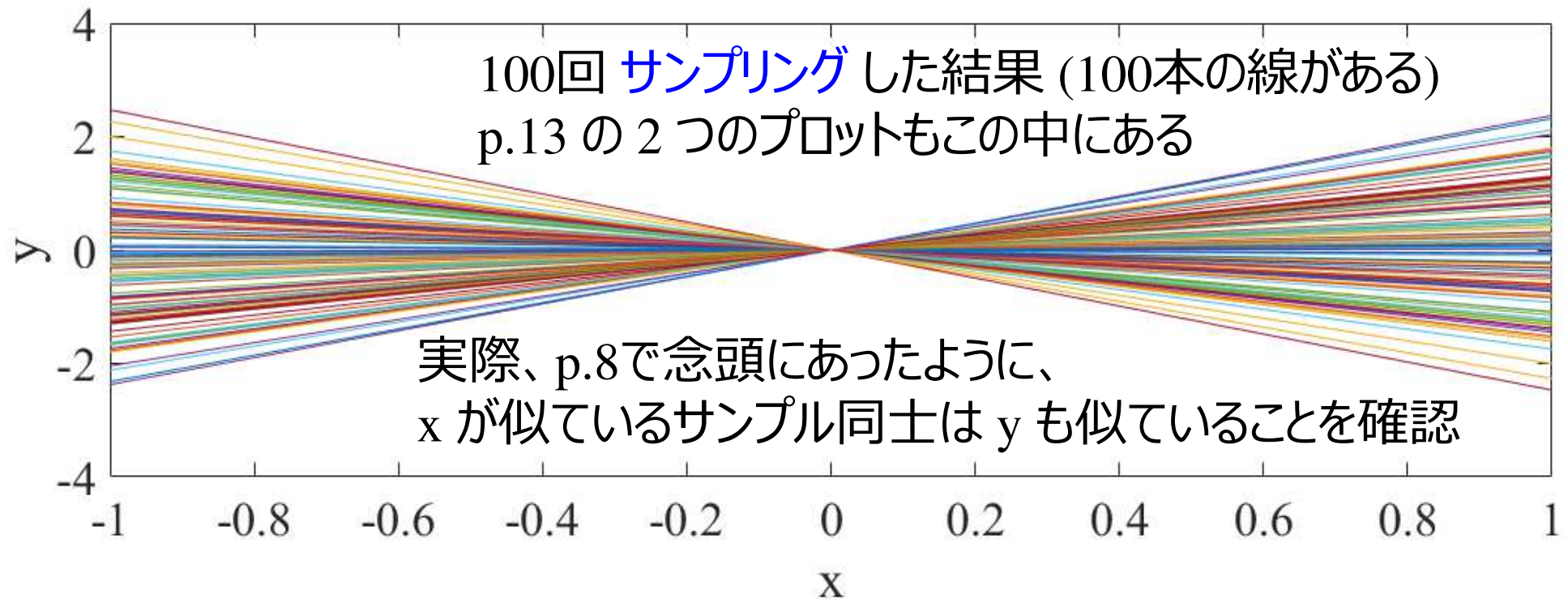
そして、すべてにおいて x と y との間をプロットし、
様子を確認する

② サンプルの結果

✓ x を、-1, -0.95, -0.9, ..., 0.9, 0.95, 1 とする

✓ $\sigma_b = 1$ とする

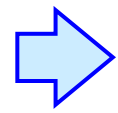
(先ほどは点で表示しましたが、今回は見やすいように線で繋いでいます)



x の値が 1 つ与えられたとき、 y の値にばらつきがある $\rightarrow y$ は分布ということ

② 説明変数の数を複数に

説明変数の数 : $1 \rightarrow m$

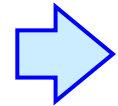


b の数 : $1 \rightarrow m$

b の分布の数 : $1 \rightarrow m$

b の分布の平均はすべて 0

b の分布の分散はすべて σ_b^2



b の分布の間の共分散はすべて 0

→ 回帰係数同士は独立しているということ

② y の平均ベクトルと分散共分散行列の計算¹⁷

✓ $y = \mathbf{x}b$ から、 i 番目のサンプルについては $y^{(i)} = \mathbf{x}^{(i)}\mathbf{b}$

✓ \mathbf{b} の平均はすべて0、分散はすべて σ_b^2 、共分散はすべて0

➡
$$m_i = E[y^{(i)}] = E[\mathbf{x}^{(i)}\mathbf{b}] = \mathbf{x}^{(i)}E[\mathbf{b}] = 0$$

$$\sigma_{y^{(i)}, y^{(j)}}^2 = \text{cov}[y^{(i)}, y^{(j)}] = \text{cov}[\mathbf{x}^{(i)}\mathbf{b}, \mathbf{x}^{(j)}\mathbf{b}]$$

$$= E[\mathbf{x}^{(i)}\mathbf{b}(\mathbf{x}^{(j)}\mathbf{b})^T] = \mathbf{x}^{(i)}E[\mathbf{b}\mathbf{b}^T]\mathbf{x}^{(j)T}$$

y の平均0より、
共分散は
内積の平均
(期待値)

$$= \mathbf{x}^{(i)}\text{cov}[\mathbf{b}, \mathbf{b}]\mathbf{x}^{(j)T} = \sigma_b^2 \mathbf{x}^{(i)}\mathbf{x}^{(j)T}$$

= \mathbf{b} の分散

$E[*]$: * の平均

$\text{cov}[* , \cdot]$: * と \cdot との間の共分散

② y の平均ベクトルと分散共分散行列 まとめ 18

$$\begin{aligned}
 \mathbf{m} &= \begin{bmatrix} m_1 \\ \vdots \\ m_i \\ \vdots \\ m_n \end{bmatrix} = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad \Sigma = \begin{bmatrix} \sigma_{y1,1}^2 & \cdots & \sigma_{y1,j}^2 & \cdots & \sigma_{y1,n}^2 \\ \vdots & & \vdots & & \vdots \\ \sigma_{yi,1}^2 & \cdots & \sigma_{yi,j}^2 & \cdots & \sigma_{yi,n}^2 \\ \vdots & & \vdots & & \vdots \\ \sigma_{yn,1}^2 & \cdots & \sigma_{yn,j}^2 & \cdots & \sigma_{yn,n}^2 \end{bmatrix} \\
 &= \sigma_b^2 \begin{bmatrix} \mathbf{X}^{(1)} \mathbf{X}^{(1)T} & \cdots & \mathbf{X}^{(1)} \mathbf{X}^{(j)T} & \cdots & \mathbf{X}^{(1)} \mathbf{X}^{(n)T} \\ \vdots & & \vdots & & \vdots \\ \mathbf{X}^{(i)} \mathbf{X}^{(1)T} & \cdots & \mathbf{X}^{(i)} \mathbf{X}^{(j)T} & \cdots & \mathbf{X}^{(i)} \mathbf{X}^{(n)T} \\ \vdots & & \vdots & & \vdots \\ \mathbf{X}^{(n)} \mathbf{X}^{(1)T} & \cdots & \mathbf{X}^{(n)} \mathbf{X}^{(j)T} & \cdots & \mathbf{X}^{(n)} \mathbf{X}^{(n)T} \end{bmatrix}
 \end{aligned}$$

③ 非線形モデルへの拡張

yの平均ベクトルと分散共分散行列で大事なものは、Xのサンプル間の

内積に b の分散をかけたもの $\sigma_{y_i, j}^2 = \sigma_b^2 \mathbf{X}^{(i)} \mathbf{X}^{(j)T}$ だけ



カーネルトリック

詳しくはこちら

<https://datachemeng.com/supportvectormachine/>

③ カーネルトリック

線形モデル (元の空間) : $y^{(i)} = \mathbf{x}^{(i)} \mathbf{b}$

↓ 高次元空間への写像 (非線形写像) : $\mathbf{x} \rightarrow \phi(\mathbf{x})$

非線形モデル関数 (高次元空間) : $y^{(i)} = \phi(\mathbf{x}^{(i)}) \mathbf{b}$

$$\sigma_{y^{(i)}, y^{(j)}}^2 = \sigma_b^2 \mathbf{x}^{(i)} \mathbf{x}^{(j)T} \quad \Rightarrow \quad \sigma_{y^{(i)}, y^{(j)}}^2 = \sigma_b^2 \phi(\mathbf{x}^{(i)}) \phi(\mathbf{x}^{(j)})^T \\ = K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$$

K : カーネル関数

③ カーネル関数の例

✓線形カーネル

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \mathbf{x}^{(i)\top} \mathbf{x}^{(j)}$$

✓ガウシアンカーネル

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \exp\left(-\frac{\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2}{2\sigma^2}\right) = \exp\left(-\gamma\|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2\right)$$

✓多項式カーネル

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \left(1 + \lambda \mathbf{x}^{(i)\top} \mathbf{x}^{(j)}\right)^d$$

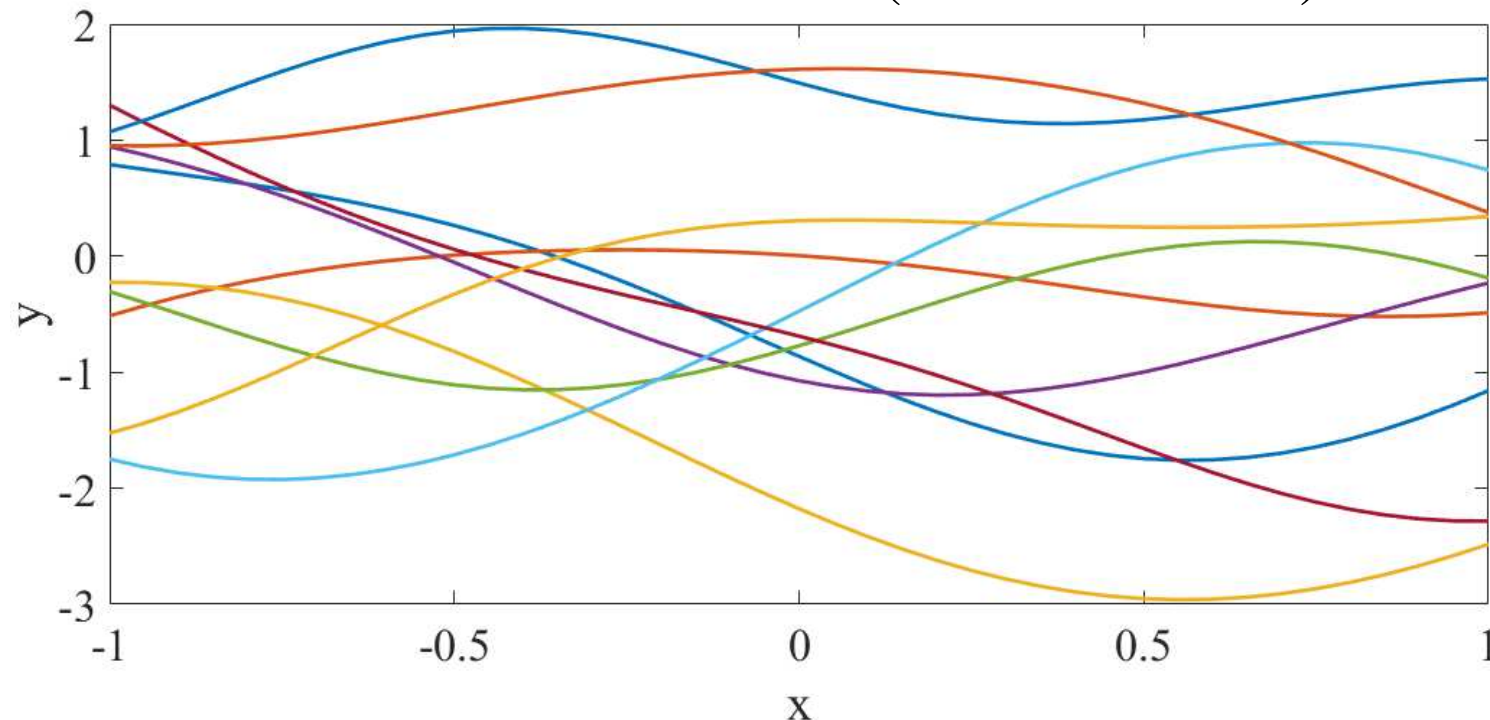
ただし、ここでは「カーネル関数で非線形性を考慮できる」といった理解で進んでいただき、GPR でよく使うカーネルやカーネルの設計については後に説明します

③ 非線形モデルのサンプリングの結果

✓ x を1変数とし、 $-1, -0.95, -0.9, \dots, 0.9, 0.95, 1$ とする

✓ ガウシアンカーネルで $\gamma = 1$ とする

10回 サンプルした結果 (10本の線がある)



p.8で念頭にあったように、 x が似ているサンプル同士は y も似ていることを確認

x の値が1つ与えられたとき、 y の値にばらつきがある $\rightarrow y$ は分布ということ

④ y に測定誤差を仮定

✓ y に測定誤差があり、その測定誤差は平均 : 0、分散 : σ_e^2 のサンプルごとに独立な正規分布に従うと仮定

$$y_{\text{obs}}^{(i)} = y^{(i)} + e^{(i)}$$

$y_{\text{obs}}^{(i)}$: 測定誤差を含む
 i 番目のサンプルの
目的変数の値

$e^{(i)}$: i 番目のサンプルの
測定誤差

$$p(e^{(i)}) = N(e^{(i)} | 0, \sigma_e^2)$$



$e^{(i)}$ の確率分布 (probability distribution) は
正規分布 (Normal distribution) であり、
平均 : 0、分散 : σ_e^2 である、という意味

④ y_{obs} の平均ベクトル

✓ p. 18 より、 $y^{(i)}$ の平均は 0

✓ $e^{(i)}$ の平均は 0

よって、 $y_{\text{obs}}^{(i)} = y^{(i)} + e^{(i)}$ より、 $y_{\text{obs}}^{(i)}$ の平均 $m_{\text{obs},i}$ も 0

④ y_{obs} の分散共分散行列

- ✓ p. 18 より、 $y^{(i)}$ と $y^{(j)}$ との間の共分散 (分散) は $\sigma_b^2 \mathbf{X}^{(i)} \mathbf{X}^{(j)T}$
 - その後、③ でカーネル関数で表したが、とりあえずカーネル関数を用いる前で考える
- ✓ $e^{(i)}$ と $e^{(j)}$ との間の共分散(分散)は、サンプルごとに独立なので、 $\delta_{i,j} \sigma_e^2$
 - $\delta_{i,j}$ は、 $i = j$ のとき 1、それ以外は 0 となる変数
 - つまり、分散が σ_e^2 で共分散が 0 ということ

よって、 $y_{\text{obs}}^{(i)} = y^{(i)} + e^{(i)}$ より、

$y^{(i)}$ と $e^{(i)}$ とが互いに独立であることから、
 $y_{\text{obs}}^{(i)}$ と $y_{\text{obs}}^{(j)}$ との間の共分散 (分散) $\sigma_{y_{\text{obs}} i, j}^2$ は、

$$\sigma_{y_{\text{obs}} i, j}^2 = \sigma_b^2 \mathbf{X}^{(i)} \mathbf{X}^{(j)T} + \delta_{i,j} \sigma_e^2$$

④ y_{obs} の分散共分散行列 まとめ

✓ サンプル数 n として、分散共分散行列を Σ_n とすると、

$$\Sigma_n = \begin{bmatrix} \sigma_b^2 \mathbf{X}^{(1)} \mathbf{X}^{(1)\text{T}} + \sigma_e^2 & \cdots & \sigma_b^2 \mathbf{X}^{(1)} \mathbf{X}^{(j)\text{T}} & \cdots & \sigma_b^2 \mathbf{X}^{(1)} \mathbf{X}^{(n)\text{T}} \\ \vdots & & \vdots & & \vdots \\ \sigma_b^2 \mathbf{X}^{(i)} \mathbf{X}^{(1)\text{T}} & \cdots & \sigma_b^2 \mathbf{X}^{(i)} \mathbf{X}^{(j)\text{T}} + \sigma_e^2 & \cdots & \sigma_b^2 \mathbf{X}^{(i)} \mathbf{X}^{(n)\text{T}} \\ \vdots & & \vdots & & \vdots \\ \sigma_b^2 \mathbf{X}^{(n)} \mathbf{X}^{(1)\text{T}} & \cdots & \sigma_b^2 \mathbf{X}^{(n)} \mathbf{X}^{(j)\text{T}} & \cdots & \sigma_b^2 \mathbf{X}^{(n)} \mathbf{X}^{(n)\text{T}} + \sigma_e^2 \end{bmatrix}$$

④ GPRのカーネル関数の特徴

- ✓ X の内積 $\mathbf{x}^{(i)}\mathbf{x}^{(j)T}$ だけでなく、 y_{obs} の分散もしくは共分散の全体をカーネル関数で表す (scikit-learn ではこの考え方)
 - p. 25, 26 より $\sigma_{y_{\text{obs}} i,j}^2 = \sigma_b^2 \mathbf{x}^{(i)}\mathbf{x}^{(j)T} + \delta_{i,j} \sigma_e^2$
 - 以下の項をカーネル関数に含める必要がある
 - σ_b^2 としての定数項の積
 - $\delta_{i,j} \sigma_e^2$ としての $i = j$ のときのみ定数項の和
- ✓ GPR では、後述するように最尤推定法でカーネル関数のパラメータを最適化できるため、比較的複雑なカーネル関数が多い

④ GPRで使われるカーネル関数の例

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \theta_0 \exp\left\{-\frac{\theta_1}{2} \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2\right\} + \theta_2$$

scikit-learn: ConstantKernel() * RBF() + WhiteKernel()

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \theta_0 \exp\left\{-\frac{\theta_1}{2} \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|^2\right\} + \theta_2 + \theta_3 \sum_{k=1}^m x_k^{(i)} x_k^{(j)}$$

scikit-learn: ConstantKernel() * RBF() + WhiteKernel() + ConstantKernel() * DotProduct()

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \theta_0 \exp\left\{-\frac{1}{2} \sum_{k=1}^m \theta_{1,k} (x_k^{(i)} - x_k^{(j)})^2\right\} + \theta_2$$

GPy の ARD
(Automatic
Relevance
Determination) に相当

scikit-learn: ConstantKernel() * RBF(np.ones(*n_features*)) + WhiteKernel()

④ GPRで使われるカーネル関数の例

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \theta_0 \exp \left\{ -\frac{1}{2} \sum_{k=1}^m \theta_{2,k} (x_k^{(i)} - x_k^{(j)})^2 \right\} + \theta_2 + \theta_3 \sum_{k=1}^m x_k^{(i)} x_k^{(j)}$$

scikit-learn: `ConstantKernel() * RBF(np.ones(n_features)) + WhiteKernel() + ConstantKernel() * DotProduct()`

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \theta_0 \left(1 + \frac{\sqrt{3}d_{i,j}}{\theta_1} \right) \exp \left(-\frac{\sqrt{3}d_{i,j}}{\theta_1} \right) + \theta_2$$

scikit-learn: `ConstantKernel() * Matern(nu=1.5) + WhiteKernel()`

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \theta_0 \left(1 + \frac{\sqrt{3}d_{i,j}}{\theta_1} \right) \exp \left(-\frac{\sqrt{3}d_{i,j}}{\theta_1} \right) + \theta_2 + \theta_3 \sum_{k=1}^m x_k^{(i)} x_k^{(j)}$$

scikit-learn: `ConstantKernel() * Matern(nu=1.5) + WhiteKernel() + ConstantKernel() * DotProduct()`

$$\text{ただし、 } d_{i,j} = \sqrt{\sum_{k=1}^m (x_k^{(i)} - x_k^{(j)})^2}$$

④ GPRで使われるカーネル関数の例

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \theta_0 \exp\left(-\frac{d_{i,j}}{\theta_1}\right) + \theta_2$$

scikit-learn: ConstantKernel() * Matern(nu=0.5) + WhiteKernel()

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \theta_0 \exp\left(-\frac{d_{i,j}}{\theta_1}\right) + \theta_2 + \theta_3 \sum_{k=1}^m x_k^{(i)} x_k^{(j)}$$

scikit-learn: ConstantKernel() * Matern(nu=0.5) + WhiteKernel() + ConstantKernel() * DotProduct()

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \theta_0 \left(1 + \frac{\sqrt{5}d_{i,j}}{\theta_1} + \frac{5d_{i,j}^2}{3\theta_1^2}\right) \exp\left(-\frac{\sqrt{5}d_{i,j}}{\theta_1}\right) + \theta_2$$

scikit-learn: ConstantKernel() * Matern(nu=2.5) + WhiteKernel()

$$K(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \theta_0 \left(1 + \frac{\sqrt{5}d_{i,j}}{\theta_1} + \frac{5d_{i,j}^2}{3\theta_1^2}\right) \exp\left(-\frac{\sqrt{5}d_{i,j}}{\theta_1}\right) + \theta_2 + \theta_3 \sum_{k=1}^m x_k^{(i)} x_k^{(j)}$$

scikit-learn: ConstantKernel() * Matern(nu=2.5) + WhiteKernel() + ConstantKernel() * DotProduct()

⑤ 問題設定

✓モデル構築用のサンプルの数を n とし、 $n+1$ 個目のサンプルにおける目的変数 y の値を推定したいとする

$$\mathbf{y}_{\text{obs}} = \begin{bmatrix} y_{\text{obs}}^{(1)} \\ \vdots \\ y_{\text{obs}}^{(i)} \\ \vdots \\ y_{\text{obs}}^{(n)} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_1^{(1)} & \cdots & x_j^{(1)} & \cdots & x_m^{(1)} \\ \vdots & & \vdots & & \vdots \\ x_1^{(i)} & \cdots & x_j^{(i)} & \cdots & x_m^{(i)} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ x_1^{(n)} & \cdots & x_j^{(n)} & \cdots & x_m^{(n)} \end{bmatrix}$$

$$\mathbf{X}_{n+1} = \begin{bmatrix} x_1^{(n+1)} & \cdots & x_j^{(n+1)} & \cdots & x_m^{(n+1)} \end{bmatrix}$$



$y_{\text{obs}}^{(n+1)}$ は？

⑤ 方針

✓ \mathbf{y}_{obs} が与えられたときの $y_{\text{obs}}^{(n+1)}$ の条件付き分布 $p(y_{\text{obs}}^{(n+1)} | \mathbf{y}_{\text{obs}})$ を求める

- これも正規分布、つまり平均と分散を求める
- これが $y_{\text{obs}}^{(n+1)}$ の予測分布、つまり平均が予測値、分散が不確実性

✓ 確率の乗法定理より、 $p(y_{\text{obs}}^{(n+1)} | \mathbf{y}_{\text{obs}})$ を求めるために、まずは同時分布 $p(\mathbf{y}_{\text{obs}}, y_{\text{obs}}^{(n+1)}) = p(\mathbf{y}_{\text{obs},n+1})$ を求める

- 同時分布とは、②でやったように y のサンプル間の分布の関係のこと (p.12参照)
- ②で求めたように、同時分布は X のサンプル間の関係で表される

$$\mathbf{y}_{\text{obs},n+1} = \begin{bmatrix} y_{\text{obs}}^{(1)} \\ \vdots \\ y_{\text{obs}}^{(i)} \\ \vdots \\ y_{\text{obs}}^{(n)} \\ y_{\text{obs}}^{(n+1)} \end{bmatrix}$$

⑤ 方針 まとめ

- ✓ $p(\mathbf{y}_{\text{obs},n+1})$ で $(n+1)$ 個のサンプル間の y のガウス分布を求める
 - $(n+1)$ 次元のガウス分布

- ✓ n 個の条件 (制約) である \mathbf{y}_{obs} により、 $(n+1) - n = 1$ 次元のガウス分布になる
 - 平均：予測値
 - 分散：予測値の不確実性

⑤ 用いる関係式

✓条件付き分布と同時分布とを結びつける式

条件付き分布 $p(\mathbf{z}_a | \mathbf{z}_b)$ の平均ベクトルを $\boldsymbol{\mu}_{a|b}$ 、
分散共分散行列を $\boldsymbol{\Sigma}_{a|b}$ とする

同時分布 $p(\mathbf{z}_a, \mathbf{z}_b)$ の平均ベクトルを $\begin{bmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{bmatrix}$

分散共分散行列を $\begin{bmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{bmatrix}$ とすると、

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab} \boldsymbol{\Sigma}_{bb}^{-1} (\mathbf{z}_b - \boldsymbol{\mu}_b)$$

詳しい導出は、

<http://www.gaussianprocess.org/gpml/chapters/RWA.pdf> のA.2

$$\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab} \boldsymbol{\Sigma}_{bb}^{-1} \boldsymbol{\Sigma}_{ba}$$

『パターン認識と機械学習 上』丸善出版 p.82-85 (第7刷)

を参照のこと

⑤ 同時分布 $p(\mathbf{y}_{\text{obs},n+1})$

✓ p. 24 より、同時分布 $p(\mathbf{y}_{\text{obs},n+1})$ の平均は $\mathbf{0}$ (0ベクトル)

✓ $p(\mathbf{y}_{\text{obs},n+1})$ の分散共分散行列を Σ_{n+1} とすると、p.25, 26より、

$$\Sigma_{n+1} = \begin{bmatrix} \Sigma_n & \mathbf{k} \\ \mathbf{k}^T & K(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+1)}) + \sigma_e^2 \end{bmatrix}$$

ただし、

$$\mathbf{k} = \left[K(\mathbf{x}^{(1)}, \mathbf{x}^{(n+1)}) \quad \dots \quad K(\mathbf{x}^{(i)}, \mathbf{x}^{(n+1)}) \quad \dots \quad K(\mathbf{x}^{(n)}, \mathbf{x}^{(n+1)}) \right]$$

⑤ 条件付き分布 $p(y_{\text{obs}}^{(n+1)} \mid Y_{\text{obs}})$

条件付き分布 $p(y_{\text{obs}}^{(n+1)} \mid \mathbf{y}_{\text{obs}})$ の

平均を $m(\mathbf{x}^{(n+1)})$ 、分散を $\sigma^2(\mathbf{x}^{(n+1)})$ とすると、p.34, 35より、

$$m(\mathbf{x}^{(n+1)}) = \mathbf{k} \boldsymbol{\Sigma}_n^{-1} \mathbf{y}_{\text{obs}}$$

$$\sigma^2(\mathbf{x}^{(n+1)}) = K(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+1)}) + \sigma_e^2 - \mathbf{k} \boldsymbol{\Sigma}_n^{-1} \mathbf{k}^T$$

GPRの使い方

✓目的変数の値を予測したいサンプルの $\mathbf{x}^{(n+1)}$ が得られたとき、

- 予測値 : $m(\mathbf{x}^{(n+1)})$
- 予測値の標準偏差 : $\sigma(\mathbf{x}^{(n+1)})$
 - 予測値が正規分布に従うと仮定すれば、 $\mathbf{x}^{(n+1)}$ の目的変数の実測値が

$m(\mathbf{x}^{(n+1)}) - \sigma(\mathbf{x}^{(n+1)}) \sim m(\mathbf{x}^{(n+1)}) + \sigma(\mathbf{x}^{(n+1)})$
の範囲に入る確率は、68.27 %

$m(\mathbf{x}^{(n+1)}) - 2 \times \sigma(\mathbf{x}^{(n+1)}) \sim m(\mathbf{x}^{(n+1)}) + 2 \times \sigma(\mathbf{x}^{(n+1)})$
の範囲に入る確率は、95.45 %

$m(\mathbf{x}^{(n+1)}) - 3 \times \sigma(\mathbf{x}^{(n+1)}) \sim m(\mathbf{x}^{(n+1)}) + 3 \times \sigma(\mathbf{x}^{(n+1)})$
の範囲に入る確率は、99.73 %

精度 β

✓ y の測定誤差の分散である σ_e^2 の代わりに、

精度 $\beta (= 1 / \sigma_e^2)$ が使われることが多い

GPRの数値例

✓モデル構築用サンプル数 $n = 3$

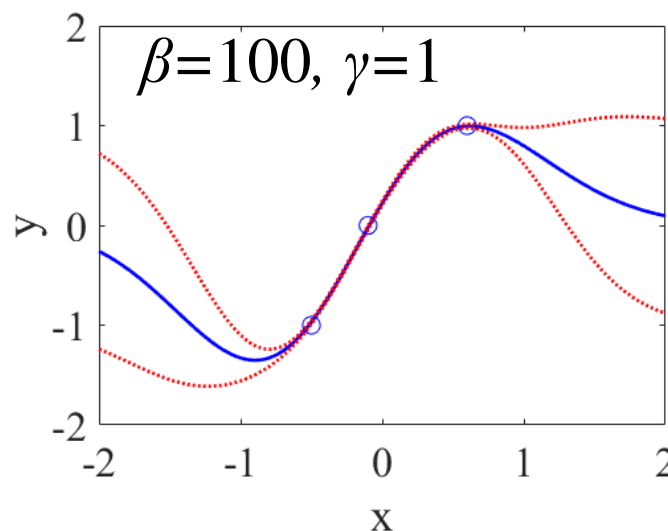
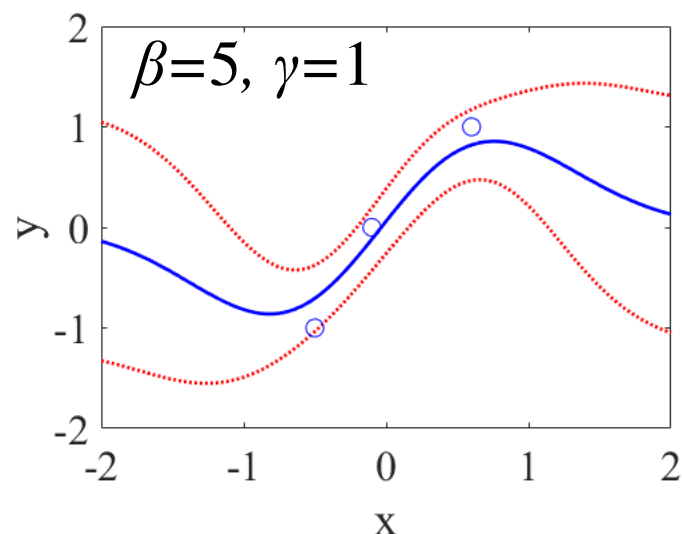
x	y
-0.5	-1
-0.1	0
0.6	1

✓予測用サンプルの x : -2, -1.99, -1.98, ..., 1.98, 1.99, 2

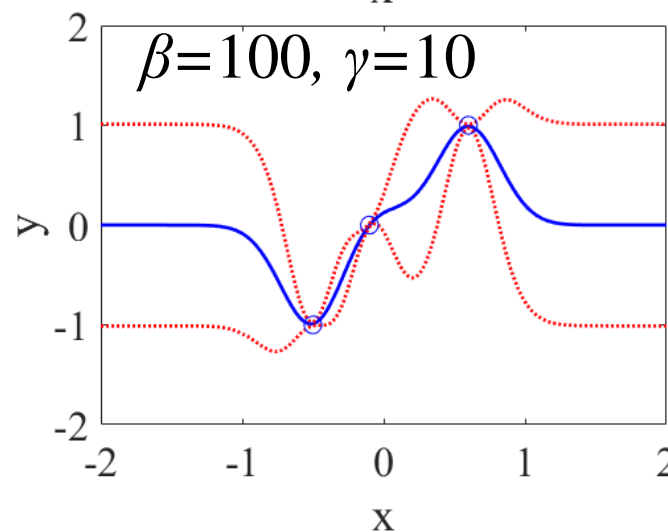
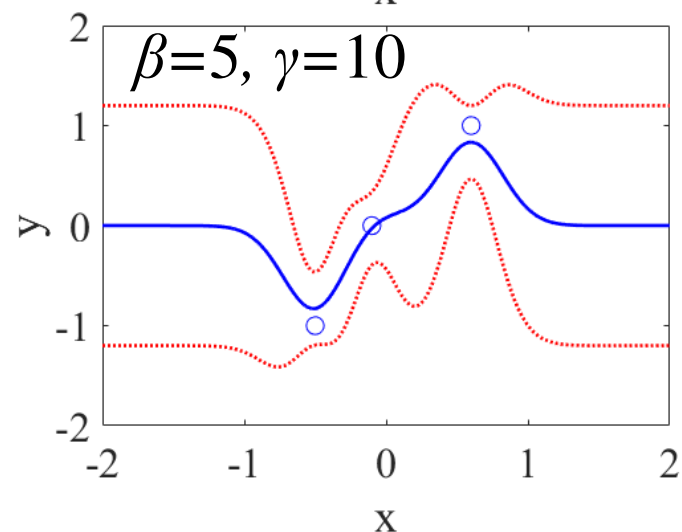
✓ガウシアンカーネル使用

GPRの数値例の結果

○ : モデル構築用サンプル、— : 予測値、⋯ : 予測値± σ



β を大きくする
(σ_e^2 を小さくする)
と、予測値が
モデル構築用
サンプルにフィット
するようになる



γ を大きくすると、
予測値や σ が
複雑な曲線に
なる

ハイパーパラメータの決め方 1/2

✓ハイパーパラメータ

- $\beta (= 1 / \sigma_e^2)$
- カーネル関数のパラメータ

✓ハイパーパラメータの決め方 3通り

- ① 事前知識から決定
 - y の測定誤差の分散が分かっているときは、それに基づいて β を設定する
 - カーネル関数のパラメータを決めることは難しいが、線形カーネルならこれでOK
- ② クロスバリデーションで最適化

ハイパーパラメータの決め方 2/2

✓ハイパーパラメータの決め方 3通り

- ③ 最尤推定・・・最も一般的な方法

- 下の対数尤度関数を最大化するパラメータベクトル θ にする

$$\ln p(\mathbf{y}_{\text{obs}} | \theta) = -\frac{1}{2} \ln |\Sigma_n| - \frac{1}{2} \mathbf{y}_{\text{obs}}^T \Sigma_n^{-1} \mathbf{y}_{\text{obs}} - \frac{n}{2} \ln(2\pi)$$

- 共役勾配法

カーネル関数の決め方

- ✓① それぞれのカーネル関数でクロスバリデーションを行い、たとえば r^2 が最も大きいカーネル関数を使用する
 - テストデータにオーバーフィットしない
 - 時間がかかる

- ✓② それぞれのカーネル関数でモデル構築し、テストデータを予測して、たとえば r^2 が最も大きいカーネル関数を使用する
 - 時間がかからない
 - テストデータにオーバーフィットするカーネル関数が選ばれる危険がある