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ない

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

✓ $\mathbf{y} = \mathbf{X}\mathbf{b}$

n：サンプル数

m：説明変数の数

$$\underset{\mathbf{y}}{\begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(i)} \\ \vdots \\ \vdots \\ y^{(n)} \end{bmatrix}} = \underset{\mathbf{X}}{\begin{bmatrix} x_1^{(1)} & \cdots & x_j^{(1)} & \cdots & x_m^{(1)} \\ \vdots & & \vdots & & \vdots \\ x_1^{(i)} & \cdots & x_j^{(i)} & \cdots & x_m^{(i)} \\ \vdots & & \vdots & & \vdots \\ \vdots & & \vdots & & \vdots \\ x_1^{(n)} & \cdots & x_j^{(n)} & \cdots & x_m^{(n)} \end{bmatrix}} \underset{\mathbf{b}}{\begin{bmatrix} b_1 \\ \vdots \\ b_j \\ \vdots \\ b_m \end{bmatrix}}$$

① 簡単にするため、まずは X を1変数とする

✓ $\mathbf{y} = \mathbf{x}b$

$\mathbf{y}$ $\mathbf{x}$ 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

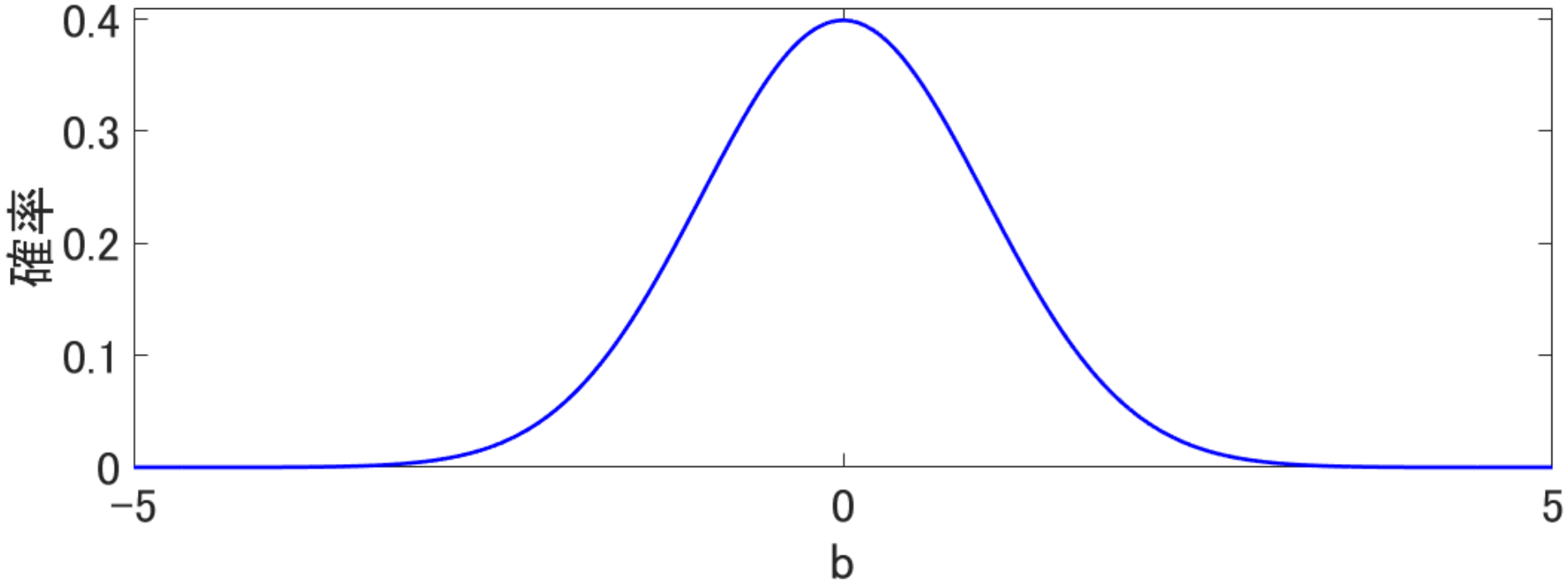
- ざっくりいうと、b は 0.1 かもしれないし、-0.4 かもしれないし、いろいろな可能性がある、ということ

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

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

✓$\sigma_b = 1$ 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 $\mathbf{\Sigma}$

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

② 平均ベクトルと分散共分散行列の計算

✓$\mathbf{y} = \mathbf{x}b$ 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 $\sigma_{\mathrm{b}}{}^2$

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

$$\sigma_{\mathrm{y}i,j}{}^2 = \mathrm{cov}\left[y^{(i)}, y^{(j)} \right] = \mathrm{cov}\left[x^{(i)}b, x^{(j)}b \right]$$

$$= x^{(i)}x^{(j)}\,\underline{\mathrm{cov}\left[b, b \right]} = x^{(i)}x^{(j)}\sigma_{\mathrm{b}}{}^2$$

$= \mathrm{b}$ の分散

E[*]：* の平均

cov[*,・]：* と・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

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

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

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

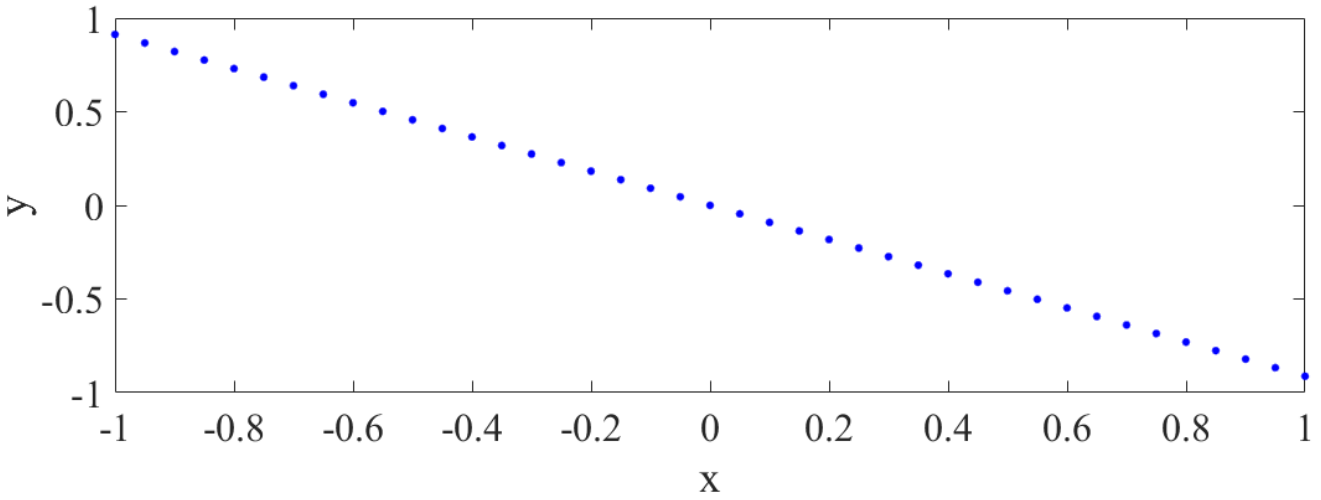
✓さらに、b の値が (分布の中から) １つに決まると、yの値が１つに決まる

② サンプルを生成してみる

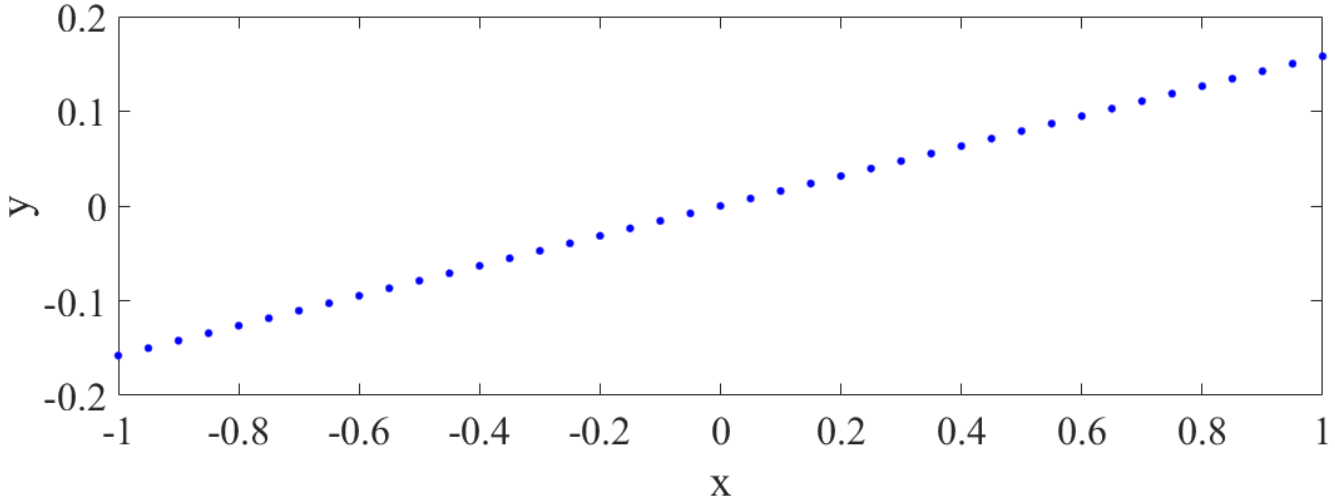
✓x を、-1, -0.95, -9, …, 0.9, 0.95, 1 とする

✓$\sigma_b = 1$ とする

b の値が 1 つに決まる



b の値が 1 つに決まる



② サンプリング

✓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
→ サンプリング

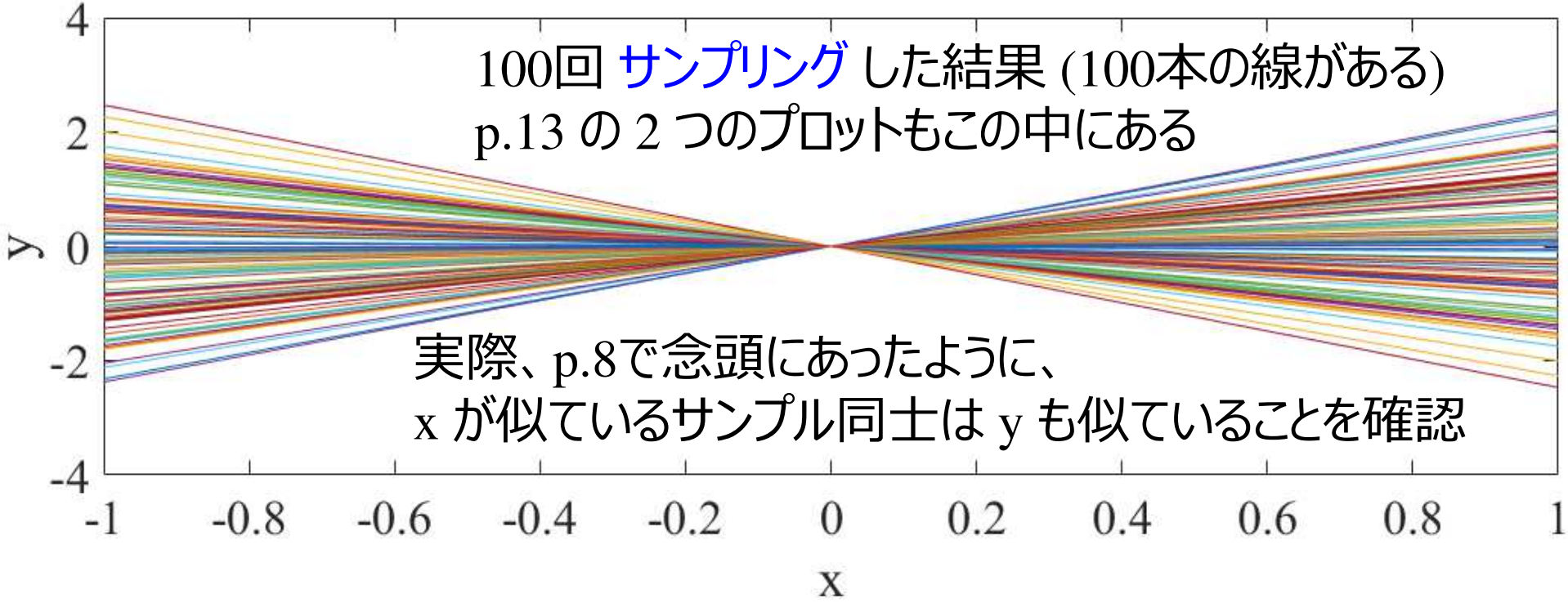
そして、すべてにおいて x と y との間の関係をプロットし、
様子を確認する

② サンプリングの結果

✓x を、-1, -0.95, -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 → 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 $\sigma_{\mathrm{b}}{}^2$

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

✓$\mathbf{y} = \mathbf{x}b$ 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

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

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

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

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

$= \mathrm{b}$ の分散

y の平均0より、
共分散は
内積の平均
(期待値)

E[*]：* の平均

cov[*,・]：* と・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

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

$$= \sigma_{\mathrm{b}}{}^2 \begin{bmatrix} \mathbf{x}^{(1)}\mathbf{x}^{(1)\mathrm{T}} & \cdots & \mathbf{x}^{(1)}\mathbf{x}^{(j)\mathrm{T}} & \cdots & \mathbf{x}^{(1)}\mathbf{x}^{(n)\mathrm{T}} \\ \vdots & & \vdots & & \vdots \\ \mathbf{x}^{(i)}\mathbf{x}^{(1)\mathrm{T}} & \cdots & \mathbf{x}^{(i)}\mathbf{x}^{(j)\mathrm{T}} & \cdots & \mathbf{x}^{(i)}\mathbf{x}^{(n)\mathrm{T}} \\ \vdots & & \vdots & & \vdots \\ \mathbf{x}^{(n)}\mathbf{x}^{(1)\mathrm{T}} & \cdots & \mathbf{x}^{(n)}\mathbf{x}^{(j)\mathrm{T}} & \cdots & \mathbf{x}^{(n)}\mathbf{x}^{(n)\mathrm{T}} \end{bmatrix}$$

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のは、Xのサンプル間の

内積にbの分散をかけたもの　$\sigma_{\mathrm{y}i,j}{}^{2} = \sigma_{\mathrm{b}}{}^{2}\mathbf{x}^{(i)}\mathbf{x}^{(j)\mathrm{T}}$　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)： $\mathbf{x} \to \phi(\mathbf{x})$

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

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

K：カーネル関数

③ カーネル関数の例

✓線形カーネル

$$K\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \mathbf{x}^{(i)}\mathbf{x}^{(j)\mathrm{T}}$$

✓ガウシアンカーネル

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

✓多項式カーネル

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

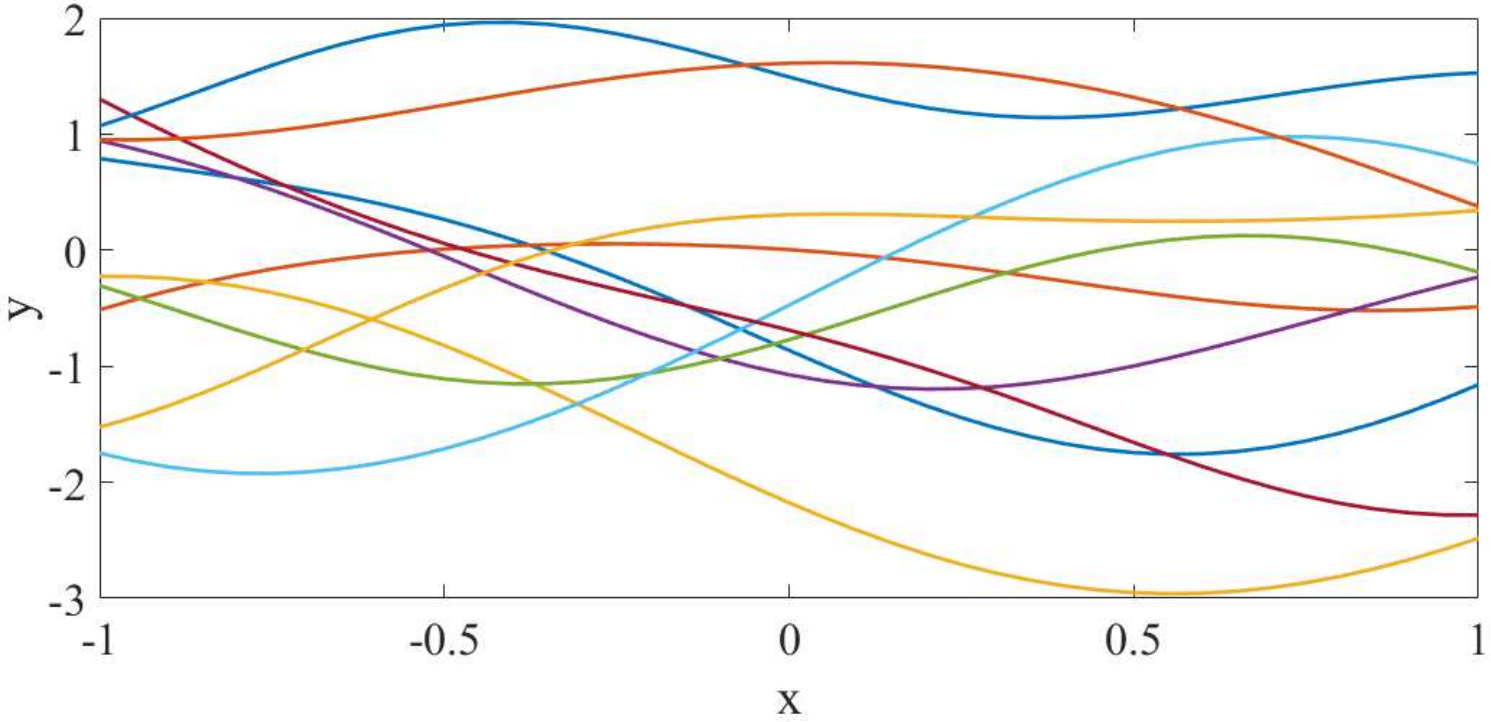
ただし、ここでは「カーネル関数で非線形性を考慮できる」といった理解で進んでいただき、GPR でよく使うカーネルやカーネルの設計については後に説明します

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, -9, …, 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 サンプリング 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 → 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

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

$y_{\mathrm{obs}}^{(i)}$：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

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

$e^{(i)}$ 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、　$y_{\mathrm{obs}}{}^{(i)} = y^{(i)} + e^{(i)}$　　より、$y_{\mathrm{obs}}{}^{(i)}$ の平均 $m_{\mathrm{obs},i}$ 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　$\sigma_{\mathrm{b}}{}^{2}\mathbf{x}^{(i)}\mathbf{x}^{(j)\mathrm{T}}$

- 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、$\delta_{i,j}\sigma_{\mathrm{e}}{}^{2}$

- $\delta_{i,j}$ は、$i=j$ のとき 1、それ以外は 0 となる変数
- つまり、分散が $\sigma_{\mathrm{e}}{}^{2}$ で共分散が 0 ということ

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

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

$$\sigma_{\mathrm{yobs}\,i,j}{}^{2} = \sigma_{\mathrm{b}}{}^{2}\mathbf{x}^{(i)}\mathbf{x}^{(j)\mathrm{T}} + \delta_{i,j}\sigma_{\mathrm{e}}{}^{2}$$

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 $\boldsymbol{\Sigma}_n$ とすると、

$$\boldsymbol{\Sigma}_n = \begin{bmatrix} \sigma_{\mathrm{b}}{}^2\mathbf{x}^{(1)}\mathbf{x}^{(1)\mathrm{T}} + \sigma_{\mathrm{e}}{}^2 & \cdots & \sigma_{\mathrm{b}}{}^2\mathbf{x}^{(1)}\mathbf{x}^{(j)\mathrm{T}} & \cdots & \sigma_{\mathrm{b}}{}^2\mathbf{x}^{(1)}\mathbf{x}^{(n)\mathrm{T}} \\ \vdots & & \vdots & & \vdots \\ \sigma_{\mathrm{b}}{}^2\mathbf{x}^{(i)}\mathbf{x}^{(1)\mathrm{T}} & \cdots & \sigma_{\mathrm{b}}{}^2\mathbf{x}^{(i)}\mathbf{x}^{(j)\mathrm{T}} + \sigma_{\mathrm{e}}{}^2 & \cdots & \sigma_{\mathrm{b}}{}^2\mathbf{x}^{(i)}\mathbf{x}^{(n)\mathrm{T}} \\ \vdots & & \vdots & & \vdots \\ \sigma_{\mathrm{b}}{}^2\mathbf{x}^{(n)}\mathbf{x}^{(1)\mathrm{T}} & \cdots & \sigma_{\mathrm{b}}{}^2\mathbf{x}^{(n)}\mathbf{x}^{(j)\mathrm{T}} & \cdots & \sigma_{\mathrm{b}}{}^2\mathbf{x}^{(n)}\mathbf{x}^{(n)\mathrm{T}} + \sigma_{\mathrm{e}}{}^2 \end{bmatrix}$$

④ GPRのカーネル関数の特徴

✓X の内積 $\mathbf{x}^{(i)}\mathbf{x}^{(j)\mathrm{T}}$ 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 $\sigma_{\mathrm{yobs}\,i,j}{}^2 = \sigma_{\mathrm{b}}{}^2\mathbf{x}^{(i)}\mathbf{x}^{(j)\mathrm{T}} + \delta_{i,j}\sigma_{\mathrm{e}}{}^2$
- 以下の項をカーネル関数に含める必要がある
  - $\sigma_{\mathrm{b}}{}^2$ としての定数項の積
  - $\delta_{i,j}\sigma_{\mathrm{e}}{}^2$ 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用いられることが多い

④ GPRで使われるカーネル関数の例

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

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

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

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

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

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

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

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

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

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

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

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

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

④ GPRで使われるカーネル関数の例

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

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

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

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

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

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

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

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

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

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

$y_{\mathrm{obs}}^{(n+1)}$ は？

⑤ 方針

✓$\mathbf{y}_{\mathrm{obs}}$ が与えられたときの $y_{\mathrm{obs}}^{(n+1)}$ の条件付き分布 $p(\, y_{\mathrm{obs}}^{(n+1)} \mid \mathbf{y}_{\mathrm{obs}} \,)$ 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 $y_{\mathrm{obs}}^{(n+1)}$ 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、$p(\, y_{\mathrm{obs}}^{(n+1)} \mid \mathbf{y}_{\mathrm{obs}} \,)$ を求めるために、まずは同時分布 $p(\, \mathbf{y}_{\mathrm{obs}} \,, y_{\mathrm{obs}}^{(n+1)} \,) = p(\, \mathbf{y}_{\mathrm{obs},n+1} \,)$ 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

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

⑤ 方針 まとめ

✓$p(\mathbf{y}_{\mathrm{obs},n+1})$ 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 $\mathbf{y}_{\mathrm{obs}}$ 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 $p(\mathbf{z}_a \mid \mathbf{z}_b)$ 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、

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

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

詳しい導出は、

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

『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($y_{obs,n+1}$)

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

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

$$\boldsymbol{\Sigma}_{n+1} = \begin{bmatrix} \boldsymbol{\Sigma}_n & \mathbf{k} \\ \mathbf{k}^{\mathrm{T}} & K\left(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+1)}\right) + \sigma_{\mathrm{e}}^{\,2} \end{bmatrix}$$

ただし、

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

⑤ 条件付き分布 p($y_{obs}^{(n+1)}$ | y_{obs})

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

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

$$\sigma^2\left(\mathbf{x}^{(n+1)}\right) = K\left(\mathbf{x}^{(n+1)}, \mathbf{x}^{(n+1)}\right) + \sigma_{\mathrm{e}}^{\,2} - \mathbf{k}\boldsymbol{\Sigma}_n^{-1}\mathbf{k}^{\mathrm{T}}$$

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 β ($= 1 / \sigma_e^2$) 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 $n = 3$

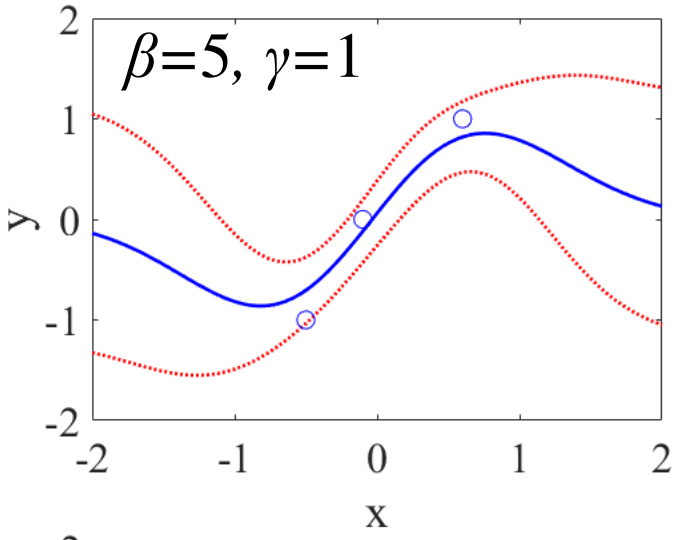
x	y
-0.5	-1
-0.1	0
0.6	1

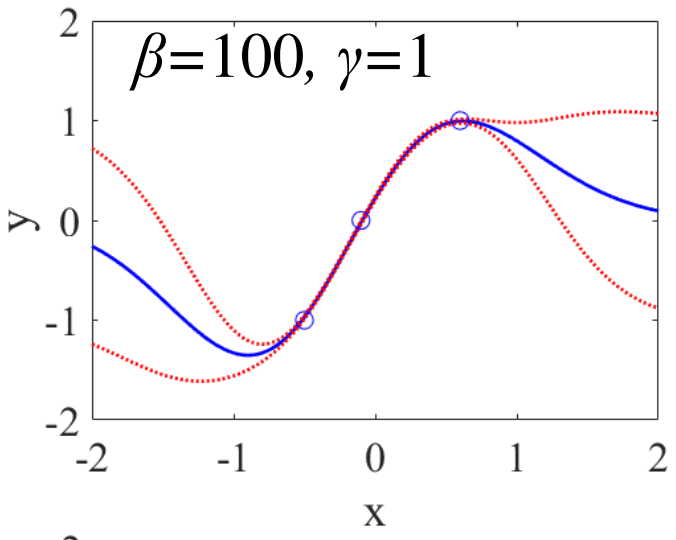
✓予測用サンプルの x：-2, -1.99, -1.98, …, 1.98, 1.99, 2

✓ガウシアンカーネル使用

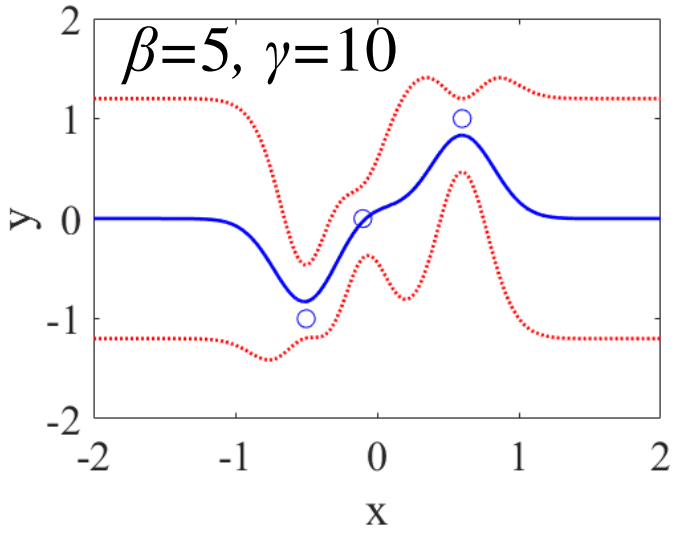
GPRの数値例の結果

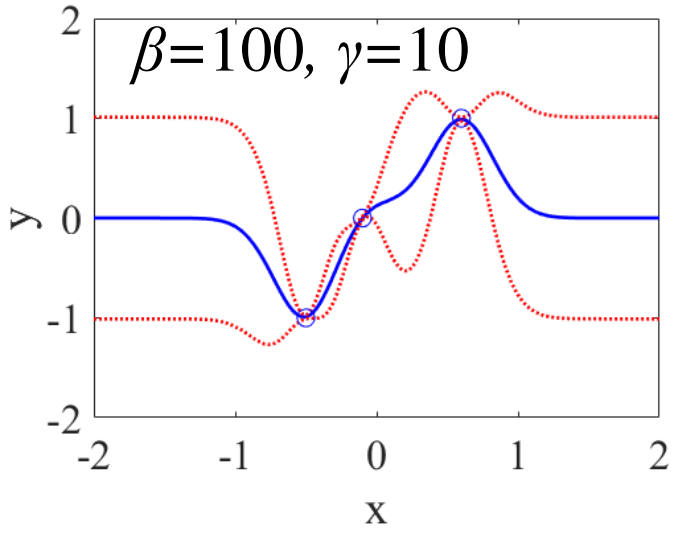
○：モデル構築用サンプル、━：予測値、・・・：予測値±σ





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

- $\beta\ (=1/\sigma_{\mathrm{e}}^{2})$
- 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 $\boldsymbol{\theta}$ にする

$$\ln p\left(\mathbf{y}_{\text{obs}} \mid \boldsymbol{\theta}\right) = -\frac{1}{2}\ln\left|\boldsymbol{\Sigma}_n\right| - \frac{1}{2}\mathbf{y}_{\text{obs}}{}^{\mathrm{T}}\boldsymbol{\Sigma}_n{}^{-1}\mathbf{y}_{\text{obs}} - \frac{n}{2}\ln\left(2\pi\right)$$

  - 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