システム/制御/情報，Vol. 62, No. 10，pp. 390–395，2018

「ガウス過程回帰の基礎から応用」特集号

解　説

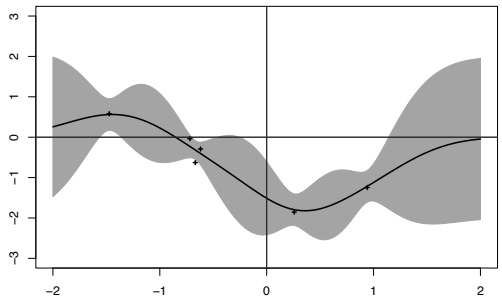
ガウス過程回帰の基礎

赤穂昭太郎*

1. はじめに

ガウス過程回帰 [2,13,11] は，入力変数 $\boldsymbol{x}$ から出力変数である実数値 y への関数 $y=f(\boldsymbol{x})$ を推定するモデルの一つである．その特徴の一つはその非線形性であり，線形回帰ではうまくフィッテイングできない場合にも有効である．もう一つ重要な特徴はベイズ推定を用いる点である．推定される関数は一つの関数ではなく，関数の分布として得られるので，推定の不確実性を表現することができる．イメージをはっきりさせるためにガウス過程回帰の例を第 1 図に示す． 6 個のデータ点に黒い曲線で示された関数でフィッティングが行われており，さらにグレーの帯はその関数の不確かさ（標準偏差）を表現している．

本稿では，ガウス過程回帰を構成するベイズ推定とガウス過程のそれぞれの基本的な部分からはじめ，ガウス過程回帰の導出やその主要な性質について解説を行う．

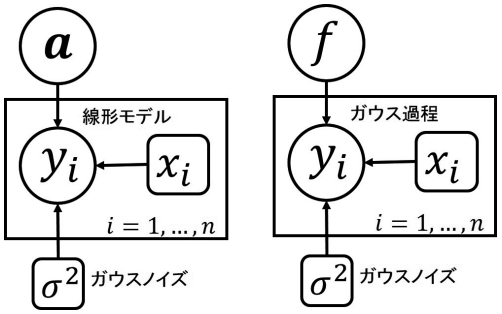


第 1 図　ガウス過程回帰の例

2. ベイズ推定

入力と出力のペア $(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_n,y_n)$ がデータとして与えられたとき，$\boldsymbol{x}$ から y への関数 $y=f(\boldsymbol{x})$ を推定することを回帰とよび，さまざまな分野で応用されている．まず回帰の単純な場合として，$\boldsymbol{x}$ は実数値ベクトルと仮定し，線形関数 $f(\boldsymbol{x})=\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}$ のパラメータ $\boldsymbol{a}$ を推定する線形回帰を考える．なお，ガウス過程回帰では入力が実数であるとか線形関数といった制限は取り払われ一般化される．

通常の線形回帰ではデータへの二乗誤差が最小になるようにパラメータ $\boldsymbol{a}$ を求めるが，ここではベイズ推定の枠組みで扱うことにする．ベイズ推定は，データを生成するプロセスを（確率的に）モデル化し，そのプロセスを逆推論する． 線形回帰の場合，まず関数パラメータ $\boldsymbol{a}$ がある確率分布 $p(\boldsymbol{a})$ に従ってランダムに生成される．そうして生成された $\boldsymbol{a}$ を用いて $\boldsymbol{x}_i$ $(i=1,\ldots,n)$ のそれぞれから $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i$ によって関数値が計算され，それにガウス分布 $N[0,\sigma^2]$ に従うノイズが加わるという確率モデル $p(y_i\,|\,\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i)$ から観測値 y_i が得られると考える．文章に書くと長くなるが，この生成プロセスをグラフィカルモデルとして記述すると第 2 図左のように単純化される．



第 2 図　回帰のグラフィカルモデル．左：線形回帰，右：ガウス過程回帰

このグラフィカルモデルをベイズの定理に基づいて逆推論することによって $\boldsymbol{a}$ の事後分布 $p(\boldsymbol{a}\,|\,D)$ (D は観測値をまとめたものを指す) が

$$p(\boldsymbol{a}\,|\,D)=\frac{1}{Z}p(\boldsymbol{a})\prod_{i=1}^{n}p(y_i\,|\,\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}_i) \tag{1}$$

によって計算され (Z は正規化定数)，新たな $\boldsymbol{x}^*$ に対する出力 y^* の予測分布 $p(y^*\,|\,D)$ は次の式で求められる．

$$p(y^*\,|\,\boldsymbol{x}^*,D)=\int p(\boldsymbol{a}\,|\,D)p(y^*\,|\,\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}^*)d\boldsymbol{a} \tag{2}$$

ガウス過程回帰では，線形関数がガウス過程 f に置き換えられるが，本質的にグラフィカルモデルの構造は同じである（第 2 図右）．すなわち，まず f がガウス過程の事前分布に従って生成され，つぎに各入力 $\boldsymbol{x}_i$ に対し

* 産業技術総合研究所 人間情報研究部門
Key Words: Kernel method, Bayesian estimation, machine learning.

て $f(\boldsymbol{x}_i)$ が計算される．その関数値に $N[0,\sigma^2]$ のガウスノイズを加えるという確率モデル $p(y_i \mid f(\boldsymbol{x}_i))$ に従って y_i が観測されるという生成モデルである．なお，とりあえず σ^2 は固定したパラメータとする．

これ以上の式の展開はとりあえず先送りにして，とりあえずガウス過程について説明することにしよう．

3. ガウス過程

ガウス過程（あるいは正規過程，Gaussian Process, 省略して GP とも）は入力空間 $\mathcal{X}$ 上のランダムな関数を定める確率過程である．確率過程というと，時間軸上に定義されているイメージがあるが，$\mathcal{X}$ は時間のような 1 次元実数空間である必要はなく，多次元ユークリッド空間でもよいし，後で述べるようにもっと一般の空間でもよい．その意味ではガウス過程という名前は誤解を招きやすくガウス確率場とでもよぶ方が適切かもしれない．

いきなり $\mathcal{X}$ が実数空間のような無限の自由度をもつ場合は理解が難しいので段階を追って説明する．まず，$\mathcal{X}$ が 1 点 x_1 だけからなるとき，1 点の関数値 $f(x_1)$ の確率分布を定めることになる (x_1 は任意の空間の要素なのであえて太字体では書かない)．$f(x_1)$ は 1 次元実数の確率変数で，複雑な確率分布でモデル化することもそれほど大変ではないが，ガウス過程では単純なガウス分布 $N[m(x_1),v(x_1,x_1)]$ でモデル化する．すなわち，$f(x_1)$ は平均 $m(x_1)$, 分散 $v(x_1,x_1)$ の二つのパラメータをもつガウス分布に従うとする．

つぎに $\mathcal{X}$ が 2 点 x_1,x_2 からなるとしよう．この場合は $f(x_1),f(x_2)$ という 2 次元の確率変数となり，2 次元ガウス分布でモデル化するとすると，平均 $(m(x_1),m(x_2))$, 分散共分散行列

$$V = \begin{pmatrix} v(x_1,x_1) & v(x_1,x_2) \\ v(x_2,x_1) & v(x_2,x_2) \end{pmatrix} \tag{3}$$

のガウス分布でモデル化できる．ただし，$v(x_2,x_1) = v(x_1,x_2)$ でなければならないので，独立なパラメータの数は 5 個（平均 2 個，分散共分散 3 個）である．

同様に $\mathcal{X}$ が $n \geq 2$ 個の点からなる場合，$\boldsymbol{f}(X_n)$ の平均と分散共分散行列がそれぞれ $\boldsymbol{m}(X_n)$, $V(X_n,X_n)$ の n 次元多変量ガウス分布でモデル化することができる．これを

$$\boldsymbol{f}(X_n) \sim N[\boldsymbol{m}(X_n),V(X_n,X_n)] \tag{4}$$

と書くことにする．ただし，式を簡明にするための記法として $X_n = (x_1,\ldots,x_n)$ とし，

$$\boldsymbol{f}(X_n) = (f(x_1),\ldots,f(x_n))^{\mathrm{T}} \tag{5}$$

$$\boldsymbol{m}(X_n) = (m(x_1),\ldots,m(x_n))^{\mathrm{T}} \tag{6}$$

$$V(X_n,X_n) = \begin{pmatrix} v(x_1,x_1) & \cdots & v(x_1,x_n) \\ \vdots & \ddots & \vdots \\ v(x_n,x_1) & \cdots & v(x_n,x_n) \end{pmatrix} \tag{7}$$

とおいた．ここで注目すべき点は，n 個に増えても，本質的には $m(x)$ という 1 変数関数と 2 個の点の間の共分散 $v(x,x')$ を定めることで確率分布が決まるということである．このように，2 点間の関係だけで分布が定まってしまうというのがガウス分布の大きな特徴である．

ガウス分布のこの性質を使えば $\mathcal{X}$ が無限の場合も同様に扱うことができる．もちろん無限個の確率変数をそのまま表現することはできないが，$\mathcal{X}$ から選んだ任意の n 個の点 X_n の同時分布が上記のガウス分布に従うとすればよい．つまり，背後には無限次元の確率変数があるが，必要に応じてそのうちの n 個の有限次元の分布として表現可能というトリックを用いて定義される．

以上をまとめて改めてガウス過程の定義を書くと，平均関数 $m(x)$, 共分散関数 $v(x,x')$ によって定義される確率過程で，任意の n 点 X_n の関数値 $\boldsymbol{f}(X_n)$ の分布が平均 $\boldsymbol{m}(X_n)$, 分散共分散行列 $V(X_n,X_n)$ の多次元ガウス分布に従うものである．

ただし 一つ注意が必要である．平均関数 $m(x)$ についてはとくに関数としての制約は必要ないが，共分散関数 $v(x,x')$ については，そこから定まる分散共分散行列が対称かつ正定値でなければならない．こうした性質をもつ関数は正定値関数とよばれ，機械学習で研究されているカーネル法で用いられるカーネル関数と同じものである [1,4].

4. ガウス過程のベイズ推定

いよいよガウス過程回帰の計算に進む．前の節で述べたように，任意の X_n に対する $\boldsymbol{f}(X_n)$ は多変量ガウス分布に従う．またガウス分布の性質から，それにガウスノイズを加えたものもガウス分布に従う．ガウス分布は周辺分布や条件付分布がやはりガウス分布として書けることから，ガウス過程回帰のすべての計算が単純な行列計算で済むことがわかる．

以下ではまず関数 f の事前分布をガウス過程として設定する．そのうえで，データが与えられたもとでの関数の事後分布をガウス過程として導出する．

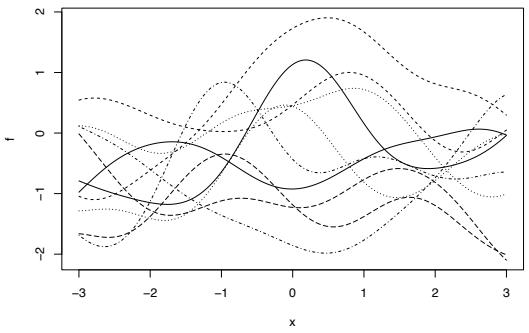
4.1 事前分布

ガウス過程の事前分布として，平均関数 $m(x) = m_0(x)$, 共分散関数 $v(x,x') = k(x,x')$ とおく．ここで，$m_0(x)$ は任意の関数であるが，とくに事前知識がなければ $m_0(x)$ は定数関数（たとえば 0）とすることが多い．$k(x,x')$ はカーネル関数であり，$\mathcal{X}$ が実数ベクトルのときの典型的なものとして，ガウスカーネル

$$k(x,x') = \exp(-\beta\|x-x'\|^2) \tag{8}$$

が挙げられる．ただし，ガウス過程だからといってガウスカーネルを使う必要はもちろんなく，さまざまなカーネル関数から適当なものを選べばよい．

ここで，具体的にガウスカーネルを共分散関数とするガウス過程の事前分布から関数をランダムにいくつか生成したのが第 3 図である．



第 3 図　ガウス過程の事前分布から生成したランダムな関数

4.2　事後分布

生成モデルをもう一度復習すると，上に述べた事前分布からランダムに f が生成され，そこから X_n に対する関数値 $\boldsymbol{f}(X_n)$ が計算され，さらに $N[\boldsymbol{0},\sigma^2 I_n]$ に従う独立なガウスノイズ $\boldsymbol{\varepsilon}$ が加わって $\boldsymbol{y}=\boldsymbol{f}(X_n)+\boldsymbol{\varepsilon}$ が観測される．

$\boldsymbol{y}$ が観測されたもとでの事後分布の平均関数と共分散関数を $\hat{m}(x)$, $\hat{v}(x,x')$ と書くことにする．これらを求めるには $\hat{m}(x)$ については $\mathcal{X}$ の 1 点上の関数値の事後分布，$\hat{v}(x,x')$ については 2 点の関数値の事後分布がわかればよいが，ここではより一般に X_n とは別の $m\geq 1$ 個の 入力点集合 $X_m=(x_1^*,\ldots,x_m^*)$ における関数値 $\boldsymbol{f}(X_m)$ の事後分布を求める．

そのために，$\boldsymbol{y}$ と $\boldsymbol{f}(X_m)$ の同時分布を考える．まず，$\boldsymbol{y}$ のもととなる $\boldsymbol{f}(X_n)$ はガウス過程の X_n における分布だから，

$$\boldsymbol{f}(X_n)\sim N[\boldsymbol{m}_0(X_n),K(X_n,X_n)] \tag{9}$$

である．ここで，$\boldsymbol{m}_0(X_n)$ は $m_0(x_i)$ を並べたベクトル，$K(X_n,X_n)$ は $k(x_i,x_j)$ を i,j 成分とする行列である．$\boldsymbol{y}$ は $\boldsymbol{f}(X_n)$ に $N[\boldsymbol{0},\sigma^2 I_n]$ のガウスノイズが加わったものだから

$$\boldsymbol{y}\sim N[\boldsymbol{m}_0(X_n),K(X_n,X_n)+\sigma^2 I_n] \tag{10}$$

となる．

一方，$\boldsymbol{f}(X_m)$ については

$$\boldsymbol{f}(X_m)\sim N[\boldsymbol{m}_0(X_m),K(X_m,X_m)] \tag{11}$$

であり，$\boldsymbol{f}(X_n)$ と $\boldsymbol{f}(X_m)$ の共分散は $K(X_n,X_m)$ である（$K(X_n,X_m)$ は $k(x_i,x_j^*)$ を i,j 成分とする行列）．

以上をまとめて書くと，$\boldsymbol{y},\boldsymbol{f}(X_m)$ の同時分布は

$$\begin{pmatrix}\boldsymbol{y}\\ \boldsymbol{f}_m\end{pmatrix}\sim N\left[\begin{pmatrix}\boldsymbol{m}_{0,n}\\ \boldsymbol{m}_{0,m}\end{pmatrix},\begin{pmatrix}K_{n,n}+\sigma^2 I_n & K_{n,m}\\ {K_{n,m}}^{\mathrm{T}} & K_{m,m}\end{pmatrix}\right] \tag{12}$$

となる．ここで式をすっきりした形にするために以下のような記法の単純化を行った．

$$\boldsymbol{f}_m=\boldsymbol{f}(X_m), \tag{13}$$
$$\boldsymbol{m}_{0,n}=\boldsymbol{m}_0(X_n),\quad \boldsymbol{m}_{0,m}=\boldsymbol{m}_0(X_m), \tag{14}$$
$$K_{n,n}=K(X_n,X_n),\quad K_{m,m}=K(X_m,X_m) \tag{15}$$
$$K_{n,m}=K(X_n,X_m)=K(X_m,X_n)^{\mathrm{T}} \tag{16}$$

あとは，$\boldsymbol{y}$ が与えられたもとで $\boldsymbol{f}_m$ の条件付分布を求めればよい．一般に，ベクトル $\boldsymbol{a}$, $\boldsymbol{b}$ が

$$\begin{pmatrix}\boldsymbol{a}\\ \boldsymbol{b}\end{pmatrix}\sim N\left[\begin{pmatrix}\boldsymbol{\mu}_a\\ \boldsymbol{\mu}_b\end{pmatrix},\begin{pmatrix}V_{aa} & V_{ab}\\ V_{ab}^{\mathrm{T}} & V_{bb}\end{pmatrix}\right] \tag{17}$$

という分布に従うとき，$\boldsymbol{a}$ が与えられたもとでの $\boldsymbol{b}$ の条件付分布を

$$\boldsymbol{b}\,|\,\boldsymbol{a}\sim N[\boldsymbol{\mu}_{b|a},V_{b|a}] \tag{18}$$

と書くことにすると，

$$\boldsymbol{\mu}_{b|a}=\boldsymbol{\mu}_b+V_{ab}^{\mathrm{T}}V_{aa}^{-1}(\boldsymbol{a}-\boldsymbol{\mu}_a), \tag{19}$$
$$V_{b|a}\quad =V_{bb}-V_{ab}^{\mathrm{T}}V_{aa}^{-1}V_{ab} \tag{20}$$

である．これらの式に (12) 式を代入すると，

$$\boldsymbol{f}_m\,|\,\boldsymbol{y}\sim N[\hat{\boldsymbol{m}},\hat{V}] \tag{21}$$
$$\hat{\boldsymbol{m}}=\boldsymbol{m}_{0,m}+K_{n,m}^{\mathrm{T}}(K_{n,n}+\sigma^2 I_n)^{-1}(\boldsymbol{y}-\boldsymbol{m}_{0,n}) \tag{22}$$
$$\hat{V}=K_{m,m}-K_{n,m}^{\mathrm{T}}(K_{n,n}+\sigma^2 I_n)^{-1}K_{n,m} \tag{23}$$

となる．

事後分布の平均関数は (22) 式のベクトルの 1 成分を取り出せばよいので

$$\hat{m}(x)=m_0(x)+\boldsymbol{k}_n(x)^{\mathrm{T}}(K_{n,n}+\sigma^2 I_n)^{-1}(\boldsymbol{y}-\boldsymbol{m}_{0,n}) \tag{24}$$

となる．ここで，$\boldsymbol{k}_n(x)=(k(x_1,x),\ldots,k(x_n,x))^{\mathrm{T}}$ とおいた．

一方，共分散関数は (23) 式の行列の 1 成分を取り出せばよく，

$$\hat{v}(x,x')=k(x,x')-\boldsymbol{k}_n(x)^{\mathrm{T}}(K_{n,n}+\sigma^2 I_n)^{-1}\boldsymbol{k}_n(x') \tag{25}$$

となる．

事後分布の平均関数で，とくに $m_0(x)$ が定数 0 のときは，

$$\hat{m}(x) = \boldsymbol{k}_n(x)^{\mathrm{T}}(K_{n,n}+\sigma^2 I_n)^{-1}\boldsymbol{y} \tag{26}$$

となるが，これはカーネル回帰で得られる関数と等しい．ガウス分布は分布の期待値とモードは一致するので，ガウス過程の期待値・モードはカーネル回帰の関数と一致することを意味する．

またガウス分布の場合，予測分布は単純で，$\boldsymbol{y}$ を観測したとき，新たな x^* に対する出力 y^* の予測分布は，単に $f(x^*)$ の分散にノイズ分散を足せばよいので，

$$y^* \mid \boldsymbol{y} \sim N[\hat{m}(x^*), \hat{v}(x^*,x^*)+\sigma^2] \tag{27}$$

となる．

予測分布を具体的に図示したのが本稿の最初に示した第 1 図 である．黒い曲線が $\hat{m}(x)$ をすべての x に対してプロットしたもので，グレーの帯はその曲線から上下方向に標準偏差 $\sqrt{\hat{v}(x,x)+\sigma^2}$ の幅で描いたプロットである．データ点の近くでは標準偏差が小さく，データから離れているところでは標準偏差が大きくなっている様子がわかる．なお，第 1 図では $\beta=1$ のガウスカーネルを用いて，ノイズ分散 $\sigma^2=0.2^2$ として求めたガウス過程回帰の結果である．

第 1 図 の表示はわかりやすいが，それだけでは分布を完全に指定することはできない．ガウス分布なので本来は 二つの x,x' における y の間の共分散の情報が必要である．y にのるノイズは x ごとに独立と仮定しているので共分散は $\hat{v}(x,x')+\sigma^2\delta(x,x')$ と書ける．ただし $\delta(x,x')$ は $x=x'$ のとき 1 で，それ以外のとき 0 とする．第 1 図に対応する共分散を示したのが第 4 図であり，x-x' の 2 次元平面上に共分散の値を等高線図で示している．第 1 図の各点の標準偏差は $x=x'$ 上の情報だけ取り出したものである． なお，第 4 図では負の値を取るところはグレーで色付けしてある．これを見ると，x, x' の組合せによって y の値に負の相関構造があることがわかる．

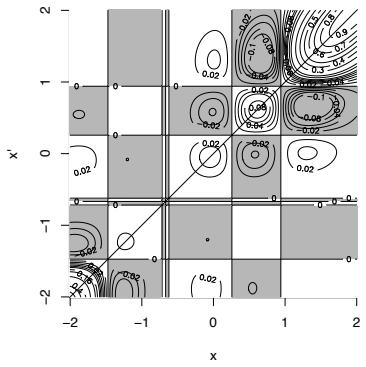
ここまででとりあえずガウス過程回帰の導出は終わりである．ここから先は，実際に計算する際の計算法の話や，関連する研究についてのあらましを紹介する．

5. 近似計算法

ガウス過程回帰は行列演算だけですべて閉じた形の計算で済むというメリットがある一方，データ数 n が増えると $n\times n$ 行列に対する計算が必要で，メモリ量も計算量も莫大なものになってしまう．

そのため，サンプル数がある程度以上大きい場合は定義通り計算するのではなく，近似計算を行うことによって現実的な時間内で計算を行う工夫がいろいろ提案されている [10,11]．基本的な計算はカーネル法と共通しているので，そちらで開発された手法はほぼそのまま使うことができる [4]．

大まかに分けると，近似計算法は以下の二つに分けられる．



第 4 図　予測値の共分散 (第 1 図と同じデータ)

- n 個あるデータの個数を減らす．
- $n\times n$ 行列 $K_{n,n}$ を低ランク行列 Q の積 $K\simeq QQ^{\mathrm{T}}$ で近似する．

前者の手法は，データ数を減らすためのさまざまな手法があるが，潜在変数を用いた統一的な説明がされている [10]．ここで，X_n や X_m とは別の入力集合 X_l を考える (X_n と重複してもよい)． $\boldsymbol{f}_n$, $\boldsymbol{f}_m$ の同時分布は

$$p(\boldsymbol{f}_n,\boldsymbol{f}_m) = \int p(\boldsymbol{f}_n,\boldsymbol{f}_m \mid \boldsymbol{f}_l)p(\boldsymbol{f}_l)d\boldsymbol{f}_l \tag{28}$$

と書けるが，右辺の条件付確率を

$$p(\boldsymbol{f}_n,\boldsymbol{f}_m \mid \boldsymbol{f}_l) \simeq p(\boldsymbol{f}_n \mid \boldsymbol{f}_l)p(\boldsymbol{f}_m \mid \boldsymbol{f}_l) \tag{29}$$

と近似する．さらに，右辺の 二つの分布をなんらかの方法で近似することによって計算量を削減する．単純なやり方の一例を挙げると，X_l は X_n の部分集合として適当に選んだうえで，$p(\boldsymbol{f}_n \mid \boldsymbol{f}_l)$ のガウス分布の分散を 0 にして決定論的な計算にしてしまうという方法がある．これによって $n\times n$ 行列を保持する必要がなくなる．このほかカーネル法で研究されている Kernel herding や random kitchen sink などとよばれる方法もデータ数を減らす手法の一種であり，ガウス過程回帰の計算に応用できる．

つぎに，低ランク行列で近似する方法について述べる．$n\times l$ 行列 Q に対して $(l<n)$，$K_{n,n}\simeq QQ^{\mathrm{T}}$ が成り立つとすると，ガウス過程回帰の計算の主要な部分を占める $(K_{n,n}+\sigma^2 I_n)^{-1}$ の計算が，

$$(K_{n,n}+\sigma^2 I_n)^{-1} \simeq \frac{1}{\sigma^2}[I_n - Q(\sigma^2 I_l + Q^{\mathrm{T}}Q)^{-1}Q^{\mathrm{T}}] \tag{30}$$

と低次元の行列の逆行列計算に帰着される．$K_{n,n}$ の分解としては不完全 Cholesky 分解や Nyström 近似などが知られている [4]．

6. モデル選択

ガウス過程回帰では，ノイズの分散 σ^2 やカーネル関数の種類，カーネル関数に含まれるパラメータなどをどう選ぶかが問題となる．そうした問題は統計学や機械学習でモデル選択の問題として盛んに研究されている．選ぶべき値は連続値の場合は，通常は有限個の候補を考え，それぞれの場合に何らかの選択規準を計算し，最も適したものを選ぶ．ここでは，数多く存在する選択規準のうち，代表的なものについて簡単に触れておく．

まず，回帰の問題でよく使われるのが交差検証法とよばれるもので，データ集合をあてはめに使う学習データと誤差を評価するテストデータに分割し，学習データで学習した関数に対するテストデータの誤差を評価する．分割の仕方を変えて何回も評価した誤差の平均値をもとにモデル選択を行う方法である．ガウス過程回帰の場合は，得られるのが関数の確率分布で扱いづらいため，期待値 $\hat{m}(x)$ に対する誤差を評価することが多い．これはカーネル法におけるカーネル回帰のモデル選択と同じである．通常は，分割の回数分の学習が必要なので多くの計算量が必要であるが，$n-1$ 個を学習データ，1 個だけをテストデータとする交差検証法 (Leave-one-out cross validation) に関しては，n 個すべてのデータを使った学習結果から計算できることが知られている [1].

一方，ガウス過程回帰のベイズ的な側面に着目したモデル選択法も存在する．その代表的なものが周辺尤度最大化である．学習データ X_n における出力 $\boldsymbol{y}$ の尤度は，$\boldsymbol{f}_n$ の分布にノイズ分散を加えた $N[\boldsymbol{m}_{0,n}, K_{nn}+\sigma^2 I_n]$ にデータの値 $\boldsymbol{y}$ を代入したものになる．通常はその対数をとって，

$$L = -\frac{1}{2}(\boldsymbol{y}-\boldsymbol{m}_{0,n})^{\mathrm{T}}(K_{nn}+\sigma^2 I_n)^{-1}(\boldsymbol{y}-\boldsymbol{m}_{0,n}) - \frac{1}{2}\log\det(K_{nn}+\sigma^2 I_n) - \frac{n}{2}\log 2\pi \quad (31)$$

となり，L を最大にするようなパラメータを選ぶことによってモデル選択を行う．

7. カーネル法との関連性

すでに見てきたように，ガウス過程回帰はカーネル法と関連が深い．カーネル法はカーネル関数の線形和 $f(x)=\sum_i \alpha_i k(x_i,x)$ で表現される関数空間（再生核ヒルベルト空間） $\mathcal{K}$ での関数のモデル化を行う（ここでの x_i はデータ点とは限らない）．

カーネル法の一つの特徴は，入力空間 $\mathcal{X}$ に対する縛りが少ないことである．もちろんユークリッド空間であれば先に挙げたガウスカーネルやそのほか多項式カーネルなどが使えるし，そうでなくてもカーネル関数 $k(x,x')$ が計算できさえすれば文字列やグラフ構造といった複雑な空間に対してもカーネル法が適用可能であり [1,4]，ガウス過程回帰も行うことができる． 本稿ではおもに回帰のみを扱ったが，パターン認識や主成分分析などそのほかのカーネル法に対してもガウス過程に拡張することができる [8,11].

ただし，実はガウス過程そのもの（第 3 図に示したようなランダムな関数）は一般に再生核ヒルベルト空間 $\mathcal{K}$ の要素ではない（関数のノルムが発散することが示される [11,13]）．ただし，(22) 式を見ればわかるように，期待値はデータ点におけるカーネル関数の和と $m_0(x)$ から計算されるので，$m_0(x)$ が $\mathcal{K}$ のもとであれば，事後分布の期待値は $\mathcal{K}$ のもととなる．このように，ランダムな関数そのものはカーネル法で扱う関数クラスをはみだしていることに注意する．

8. いくつかのホットトピックス

8.1 ベイズ最適化

機械学習においてガウス過程回帰が最近注目度を増している理由に，ベイズ最適化 [3] とよばれる実験計画の最適化手法に使われていることがあげられる．とくに製造業などで，材料や工程を最適化してできるだけ性能が高い製品を作りたいという問題は常々起きている．その際，1 回 1 回の実験やシミュレーションに時間がかかることも多く，できるだけ実験回数を減らすようにしたいというニーズが強い．とくに近年マテリアルインフォマティクス [7,12] などへの応用の研究が進展している．

ベイズ最適化法の基本的な仕組みは，以下の手続きの繰り返しによって逐次実験条件の設計を行う手法である．

(1) これまで行った実験結果からガウス過程回帰を用いて応答関数を推定する．
(2) これまでの最適値より良くなる可能性の度合いを表す関数（獲得関数という）を評価する．
(3) 獲得関数が最大になるような実験設定を探索する．

獲得関数には，これまでの最適値より良くなる確率 (probability of improvement) や，良くなった場合の期待値 (expectation of improvement) などいろいろ提案されているが，いずれにしてもガウス過程回帰のように関数を分布で推定しているという性質が利用される．

8.2 入力へのノイズの影響

ガウス過程回帰をはじめ，通常の回帰では出力変数 y に対するノイズのみを考えることが多い．ただ，実際問題では入力変数にも不確実性やノイズがのることもある．この問題は統計学では昔から難しい問題として知られているが，ベイズ的にモデル化することによって解決する試みがなされている [5]．入力点にノイズがのることによって，大きく二つの問題が生じる．一つ目はもはや行列計算だけの閉じた形では分布が書き表せなくなること．もう一つは，ガウス過程回帰では実質的に有限個の入力点を扱えばよかったが，入力点にノイズがのるとすると無限の入力点を扱う必要が出てくることである．マルコフ連鎖モンテカルロ法を使ったトリッキーな手法によってこの問題を解決した研究がなされている [5].

8.3 ガウス過程間のダイバージェンス

確率分布間の隔たりを測るのに Kullback-Leibler ダイバージェンス

$$D(p,q) = E_p[\log p(X) - \log q(X)] \tag{32}$$

がよく用いられる尺度であるが，通常確率変数 X は有限次元で定義される．確率変数が無限次元のガウス過程においてはどのように定義すればよいだろうか．ガウス過程回帰で行ったように，任意の有限個の点集合 X_m を用意して，その点集合に対する分布を考えると，それは有限次元のガウス分布の間のダイバージェンスであり，二つの確率分布 $p_1(X_m) = N[\boldsymbol{m}_1(X_m), V_1(X_m,X_m)]$, $p_2(X_m) = N[\boldsymbol{m}_2(X_m), V_2(X_m,X_m)]$ の間の距離として

$$D(p_1,p_2) = \frac{1}{2}\log\det(V_2V_1^{-1}) + \mathrm{tr}(V_1^{-1}(\boldsymbol{m}_1-\boldsymbol{m}_2)(\boldsymbol{m}_1-\boldsymbol{m}_2)^{\mathrm{T}} + V_1 - V_2) \tag{33}$$

と計算される．X_m の取り方（個数や位置）をさまざまに変えたときの $D(p_1,p_2)$ の上限値を二つのガウス過程のダイバージェンスの定義としよう．

実は，事前分布が共通のガウス過程回帰があったとき，上記のダイバージェンスは $X_m = X_{n_1} \cup X_{n_2}$ としたときの上記の有限次元ガウス分布のダイバージェンスに一致することがわかっている [9,6]．ここで，X_{n_1} と X_{n_2} はそれぞれのガウス過程回帰の学習データ集合である．

これを利用して，機械学習での変分ベイズ法への応用 [9] や，ガウス過程回帰の次元縮約問題 [6] などが研究されている．ただし，事前分布が異なる場合には上記の定義ではダイバージェンスは発散してしまうので注意が必要である．

9. おわりに

ガウス過程回帰の基礎的な解説を中心に，カーネル法との関連性やベイズ最適化など最近のホットなトピックスについてもあらましを紹介した（詳細は文献参照）．本稿が読者の研究の一助になれば幸いである．

(2018 年 5 月 1 日受付)

参 考 文 献

[1] 赤穂: カーネル多変量解析: 非線形データ解析の新しい展開, 岩波書店 (2008)

[2] C. M. Bishop（元田ほか（監訳））: パターン認識と機械学習 下. ベイズ理論による統計的予測, 丸善出版 (2008)

[3] E. Brochu, V. M. Cora and N. D. Freitas: A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning; arXiv:1012.2599 (2010)

[4] 福水: カーネル法入門—正定値カーネルによるデータ解析. 朝倉書店 (2010)

[5] Y. Iba and S. Akaho: Gaussian process regression with measurement error; *IEICE Transactions on Information and Systems*, Vol. 93, No. 10, pp. 2680–2689 (2010)

[6] 石橋, 古川, 赤穂: 事後分布推定されたガウス過程間の KL ダイバージェンスは有限次元の正規分布間の KL ダイバージェンスで評価できる (情報論的学習理論と機械学習); 電子情報通信学会技術研究報告: 信学技報, Vol. 117, No. 293, pp. 155–160 (2017)

[7] S. Ju, T. Shiga, L. Feng, Z. Hou, K. Tsuda and J. Shiomi: Designing nanostructures for phonon transport via Bayesian optimization; *Physical Review X*, Vol. 7, No. 2, 021024 (2017)

[8] N. D. Lawrence: Gaussian process latent variable models for visualisation of high dimensional data; *Advances in Neural Information Processing Systems*, pp. 329–336 (2004)

[9] A. G. D. G. Matthews, J. Hensman, R. Turner and Z. Ghahramani: On sparse variational methods and the Kullback-Leibler divergence between stochastic processes; *Journal of Machine Learning Research*, Vol. 51, pp. 231–239 (2016)

[10] J. Quiñonero-Candela and C. E. Rasmussen: A unifying view of sparse approximate Gaussian process regression; *Journal of Machine Learning Research*, Vol. 6, pp. 1939–1959 (2005)

[11] C. E. Rasmussen: Gaussian processes in machine learning; *Advanced Lectures on Machine Learning*, Springer (2004)

[12] T. Ueno, H. Hino, A. Hashimoto, Y. Takeichi, M. Sawada and K. Ono: Adaptive design of an X-ray magnetic circular dichroism spectroscopy experiment with Gaussian process modelling; *NPJ Computational Materials*, Vol. 4, No. 1, p. 4 (2018)

[13] G. Wahba: *Spline Models for Observational Data*, Vol. 59, SIAM (1990)

著 者 略 歴

赤穂 昭太郎（あかほ しょうたろう）

1965 年 7 月 30 日生．1990 年 3 月東京大学大学院工学研究科計数工学専攻修士課程修了．同年 4 月電子技術総合研究所研究官，2001 年 4 月独法化に伴い産業技術総合研究所研究グループ長となり現在に至る．機械学習の研究に従事．博士（工学）．電子情報通信学会，日本神経回路学会などの会員．