＜論 説＞

# AI手法（ガウス過程）を用いた予測――理論篇

荒 木 勝 啓

目　　次



## Ⅰ　パラメトリック予測の方法と問題点

### (i)　はじめに

前著[10]において、時系列データを使い、「パラメトリック予測」と呼ばれるカテゴリーに含まれる短期予測を行った。パラメトリック予測と呼ばれる理由は、予測に使う推定量が漠然とした「関数」ではなく、具体的な関数形のパラメーターあるいはその確率的分布だからである。例えば、推定関数が単に関数である、というのではなく

$$y = ax + b$$

と仮定してパラメーター$a$の推定値を$a = 0.5$、$b$の推定値を$b = 3$などと推定するのがパラメトリック推定である。

本稿の目的は、パラメトリック予測とガウス過程と呼ばれるノンパラメトリック予測を、経済時系列データを用いて比較するため、パラメトリック予測の方法論と問題点、およびガウス過程の内容を概観することである。データの計算や図表の詳細は次稿以降に回すが、重要な公式（(60)から(65)式）の証明は、通常の説明よりも細かく行う[1]。

以下$m$次元インプット・ベクトル

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_m \end{bmatrix} \tag{1}$$

は、$m$種類の変数、例えば所得、価格、年齢などを表すデータ変数ベクトルとする。計量経済学の文脈では、$\boldsymbol{x}$は説明変数である。$n$個のインプット・データは統計の用語では対となる観測値

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

1　これらの公式はガウス過程を理解する鍵となる式であり、様々な文献や、バイブルとも呼ばれる名著Gaussian Processes for Machine Learning [4]の中でも示されているが、導出過程が若干分かりづらいので本稿では詳解する。

の第$i$番目のデータ$y_i$とともに、$(\boldsymbol{x}_i, y_i)$を一組として「$n$組のサンプル（標本）」と呼ばれるが、機械学習（Machine Learning）の文脈では$\boldsymbol{x}_i\,(i = 1,2,\cdots,n)$は学習のための訓練データと解釈され、$\boldsymbol{y}$は教師である。$n$個の訓練データ$\boldsymbol{x}_i$は$m \times n$行列

$$\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n] = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

（$m \times n$行列）

にまとめられる。$\boldsymbol{X}$はデザイン行列、または計画行列と呼ばれることもある[2]。また$\boldsymbol{X}$と$\boldsymbol{y}$をまとめてデータ$(D)$という意味で

$$D = (\boldsymbol{X}, \boldsymbol{y})$$

と書く。

多重線形回帰モデルは、定数項を$x_{11} = 1$, $x_{12} = 1, \cdots, x_{1n} = 1$、すなわち$\boldsymbol{X}$の第1行をすべて1とし、ノイズまたは攪乱項（確率モデルでない場合は「残差」）を$\varepsilon_i$、係数パラメーターを

$$\boldsymbol{w} = \begin{bmatrix} w_1 \\ \vdots \\ w_m \end{bmatrix}$$

として

$$y_i = w_1 x_{1i} + w_2 x_{2i} + \cdots + w_m x_{mi} + \varepsilon_i (i = 1,2,\cdots n)$$

$$y_i = \boldsymbol{x}_i^T \boldsymbol{w} + \varepsilon_i (i = 1,2,\cdots,n) \qquad (2)$$

あるいは$\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots \varepsilon_n)^T$として

$$\boldsymbol{y} = \boldsymbol{X}^T \boldsymbol{w} + \boldsymbol{\varepsilon} \qquad (3)$$

のように表される。

時系列分析で「期」$t_i$およびその$M$-1次までの多項式が説明変数となる非線形回帰式は

$$\boldsymbol{y}_i = w_1 + w_2 t_i + w_3 t_i^2 + \cdots + w_M t_i^{M-1} + \varepsilon_i (i = 1,2,\cdots,n) \qquad (4)$$

と表すことができる。$\varphi_1(t) = 1$, $\varphi_2(t) = t$, $\varphi_3(t) = t^2, \cdots, \varphi_M(t) = t^{M-1}$と置いて

$$\boldsymbol{\varphi}(t) = \begin{bmatrix} \varphi_1(t) \\ \vdots \\ \varphi_M(t) \end{bmatrix} = \begin{bmatrix} 1 \\ \vdots \\ t^{M-1} \end{bmatrix}$$

と定義する。より大きな次元$N\,(N \geq M)$の特徴空間の中で

$$\begin{bmatrix} \varphi_1 \\ 0 \\ 0 \\ \vdots \\ \vdots \end{bmatrix}, \begin{bmatrix} 0 \\ \varphi_2 \\ 0 \\ \vdots \\ \vdots \end{bmatrix}, \cdots, \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \varphi_M \\ \vdots \end{bmatrix}$$

が1次独立ならば、これを基底として特徴空間の中の$M$次元部分空間を考えることができる。(4)はこの部分空間における線形結合

$$y_i = \varphi_1(t_i) w_1 + \varphi_2(t_i) w_2 + \cdots + \varphi_M(t_i) w_M + \varepsilon_i$$
$$= \boldsymbol{\varphi}(t_i)^T \boldsymbol{w} + \varepsilon_i (i = 1,2,\cdots,n) \qquad (5)$$

あるいは

$$\boldsymbol{\Phi} = \left[\boldsymbol{\varphi}(t_1), \quad \boldsymbol{\varphi}(t_2), \quad \cdots, \quad \boldsymbol{\varphi}(t_n)\right]$$
$$= \begin{bmatrix} 1 & 1 & \cdots & 1 \\ t_1 & t_2 & \cdots & t_n \\ \vdots & \vdots & \cdots & \vdots \\ t_1^{M-1} & t_2^{M-1} & \cdots & t_n^{M-1} \end{bmatrix} \quad (M \times n\text{行列})$$

と置いて

$$\boldsymbol{y} = \boldsymbol{\Phi}^T \boldsymbol{w} + \boldsymbol{\varepsilon} \qquad (6)$$

と解釈することができる。すなわち(4)は1次元$(t)$空間上では非線形関数であっても特徴空間の中への写像（into mapping）(5)によって、特徴空間の中に含まれる部分空間上の線形関数へと変わる。

(4)(5)は$\boldsymbol{x} = t$という特殊ケースであるが、一般的に書くと、(1)のような$m$次元ベクトル$\boldsymbol{x}$が特徴空間の中に$M$次元ベクトルとして$\boldsymbol{\varphi}(\boldsymbol{x})$

2　Rasmussen,C.E. and Williams, K.I., [4], p.8.

のように写像され、写像された $n$ 個の $M$ 次元ベクトル $\boldsymbol{\varphi}(\boldsymbol{x}_i)(i=1,2,\cdots,n)$ を

$$\boldsymbol{\Phi}=\left[\boldsymbol{\varphi}(\boldsymbol{x}_1),\quad \boldsymbol{\varphi}(\boldsymbol{x}_2),\quad \cdots,\quad \boldsymbol{\varphi}(\boldsymbol{x}_n)\right]=\begin{bmatrix}\varphi_1(\boldsymbol{x}_1) & \cdots & \varphi_1(\boldsymbol{x}_n)\\ \vdots & \ddots & \vdots\\ \varphi_M(\boldsymbol{x}_1) & \cdots & \varphi_M(\boldsymbol{x}_n)\end{bmatrix}$$

と行列化すると、元の $m$ 次元空間上の非線形モデル

$$y_i=w_1\varphi_1(\boldsymbol{x}_i)+w_2\varphi_2(\boldsymbol{x}_i)+\cdots+w_M\varphi_M(\boldsymbol{x}_i)+\varepsilon_i(i=1,2,\cdots n)$$

が $M$ 次元部分空間上の線形式として

$$y_i==\boldsymbol{\varphi}(\boldsymbol{x}_i)^T\boldsymbol{w}+\varepsilon_i(i=1,2,\cdots n)\qquad(7)$$

および

$$\boldsymbol{y}=\boldsymbol{\Phi}^T\boldsymbol{w}+\boldsymbol{\varepsilon}\qquad(8)$$

の形に変換される。ただし $\boldsymbol{w}$ の要素数は部分空間の次元数 $M$ で、$\boldsymbol{w}=(w_1,\cdots,w_M)^T$ である。(7) は (2) の $\boldsymbol{x}_i$ を $\boldsymbol{\varphi}(\boldsymbol{x}_i)$ に、(8) は (3) の $\boldsymbol{X}$ を $\boldsymbol{\Phi}$ に置き換えたものである。また $\boldsymbol{\varphi}(\boldsymbol{x})=\boldsymbol{x}$, $M=m$ と置けば (2) 式は (7) 式の、(3) 式は (8) 式の特殊ケースであると考えることができる。

パラメトリック・非ベイズ回帰モデルの場合は $\boldsymbol{w}$ の推定量 $\hat{\boldsymbol{w}}$ が最小二乗法や最尤法により直接求められる。パラメトリック・ベイズ推定の場合は $\hat{\boldsymbol{w}}$ そのものというより $\boldsymbol{w}$ についての事前的確率分布 $p(\boldsymbol{w})$ が、訓練データ $D$ によって事後的に最適修正され（MAP（Maximum a Posteriori）推定）、$w$ の事後分布 $p(w\mid D)$ として求められるという形式をとる。その場合 $\hat{\boldsymbol{w}}$ に相当するものは事後分布の平均すなわち $\bar{\boldsymbol{w}}_{\boldsymbol{D}}=E(\boldsymbol{w}\mid D)$ である。

(7) (8) のモデルは $\boldsymbol{w}$ を係数とする 1 次結合であるから、係数 $\boldsymbol{w}$ の要素の個数 $M$ が $\boldsymbol{y}$ の次元を定める。$M$ は 1 つ 2 つと数えられる数（可算数）なので、$\boldsymbol{y}$ の次元は有限またはたかだか可算次元である。後述のようにこの限界を取り払うことがガウス過程の 1 つの意義である。

### (ii)　パラメトリック・非ベイズ推定

非ベイズ多重線形回帰モデル (3) の場合、$\boldsymbol{w}$ の最小二乗推定量はよく知られているように、データ $D$ の要素のみで

$$\hat{\boldsymbol{w}}=\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{y}\qquad(9)$$

と計算され、$\boldsymbol{y}$ の推定量は

$$\hat{\boldsymbol{y}}=\boldsymbol{X}^T\hat{\boldsymbol{w}}=\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{X}^T\right)^{-1}\boldsymbol{X}\boldsymbol{y}\qquad(10)$$

と計算される。

(8) 式の場合は $\boldsymbol{X}$ を $\boldsymbol{\Phi}$ に置き換えることによって推定量が

$$\hat{\boldsymbol{w}}=\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T\right)^{-1}\boldsymbol{\Phi}\boldsymbol{y}\qquad(11)$$

$$\hat{\boldsymbol{y}}=\boldsymbol{\Phi}^T\hat{\boldsymbol{w}}=\boldsymbol{\Phi}^T\left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T\right)^{-1}\boldsymbol{\Phi}\boldsymbol{y}\qquad(12)$$

と表される。(9) の $\boldsymbol{X}\boldsymbol{X}^T$ は

$$\boldsymbol{X}\boldsymbol{X}^T=\begin{bmatrix}\sum_{j=1}^{n}x_{1j}x_{1j} & \cdots & \sum_{j=1}^{n}x_{1j}x_{mj}\\ \vdots & \vdots & \vdots\\ \sum_{j=1}^{n}x_{mj}x_{1j} & \cdots & \sum_{j=1}^{n}x_{mj}x_{mj}\end{bmatrix}\quad(m\times m\text{ 行列})$$

$$=\begin{bmatrix}\sum_{j=1}^{n}(x_{1j})^2 & \cdots & \sum_{j=1}^{n}x_{1j}x_{mj}\\ \vdots & \vdots & \vdots\\ \sum_{j=1}^{n}x_{mj}x_{1j} & \cdots & \sum_{j=1}^{n}(x_{mj})^2\end{bmatrix}\quad(m\times m\text{ 行列})\qquad(13)$$

(11) の $\boldsymbol{\Phi}\boldsymbol{\Phi}^T$ は

$$\boldsymbol{\Phi}\boldsymbol{\Phi}^T=\begin{bmatrix}\sum_{j=1}^{n}\varphi_1(\boldsymbol{x}_j)\varphi_1(\boldsymbol{x}_j) & \cdots & \sum_{j=1}^{n}\varphi_1(\boldsymbol{x}_j)\varphi_M(\boldsymbol{x}_j)\\ \vdots & \vdots & \vdots\\ \sum_{j=1}^{n}\varphi_M(\boldsymbol{x}_j)\varphi_1(\boldsymbol{x}_j) & \cdots & \sum_{j=1}^{n}\varphi_M(\boldsymbol{x}_j)\varphi_M(\boldsymbol{x}_j)\end{bmatrix}$$

$(M\times M\text{行列})\qquad(14)$

と表され、両者ともに対称行列（要素が実数なら半正定値行列）である。対称行列として他に

代表的なものは改めて $\boldsymbol{x}_i \equiv \boldsymbol{x}_i - E(\boldsymbol{x}_i)$ と置いて表した共分散行列

$$\Sigma = \begin{bmatrix} E(\boldsymbol{x}_1^T \boldsymbol{x}_1) & \cdots & E(\boldsymbol{x}_1^T \boldsymbol{x}_n) \\ \vdots & \vdots & \vdots \\ E(\boldsymbol{x}_n^T \boldsymbol{x}_1) & \cdots & E(\boldsymbol{x}_n^T \boldsymbol{x}_n) \end{bmatrix} \quad (n \times n \text{行列})$$

がある。期待値関数 $E$ をさらに一般化して

$$k(\boldsymbol{x}', \boldsymbol{x}) = k(\boldsymbol{x}, \boldsymbol{x}')$$

という対称性を持つ関数（カーネル関数 kernel function または共分散関数とも呼ばれる）を要素に持つ行列（カーネル）

$$K = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \vdots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix} \quad (n \times n \text{行列}) \quad (15)$$

を定義すればより一般的なモデル構築が可能になるであろう[3]。これがカーネル法であり[4]、後述のガウス過程でもパラメーターに依存しないモデル（ノンパラメトリック・モデル）を構築するための基本ツールとなっている。

なお (5) 式の場合 $\boldsymbol{\Phi}\boldsymbol{\Phi}^T$ は

$$\boldsymbol{\Phi}\boldsymbol{\Phi}^T = \begin{bmatrix} n & \cdots & \sum_{i=1}^{n} t_i^{m-1} \\ \vdots & \vdots & \vdots \\ \sum_{i=1}^{n} t_i^{m-1} & \cdots & \sum_{i=1}^{n} t_i^{2(m-1)} \end{bmatrix}$$

である。株式相場の日毎データの場合1年で約250 サンプル数あり、例えば $m = 10$ とすると $\boldsymbol{\Phi}\boldsymbol{\Phi}^T$ の末項の計算量は $250^{18}$ という天文学的オーダーになり。さらに $(\boldsymbol{\Phi}\boldsymbol{\Phi}^T)^{-1}$ の計算は巨大な数の逆数（ほぼ 0 に近い）を含むので、正確な計算が困難となる[5]。パラメーターを増やすことが「次元の呪い」と呼ばれる所以であり、ノンパラメトリック・モデルを選好する 1 つの理由として挙げられる。[6]

### (iii)　パラメトリック・ベイズ推定

ベイズ推定は、推定すべき対象（パラメーターや関数など）に関する仮定（$H$）の事前確率または確率密度[7] $p(H)$ を、{その仮定のもとでデータが得られる確率: $p(D \mid H)$} / { データ（$D$）が得られる確率: エビデンス $p(D)$} で修正し、すなわち「データが得られるとしてそのうち仮定を満たすデータがどのくらいあるのかの割合: $p(D \mid H) / p(D)$ 」で修正し、実際にデータが得られた後の事後確率 $p(H \mid D)$ を

$$p(H \mid D) = \left( p(D \mid H) / p(D) \right) \times p(H) \quad (16)$$

のように修正するという、ベイズの定理に基づく推定法である。(16) 式右辺の分子の $p(D \mid H)$ は尤度（likelihood）と呼ばれ、計算は容易であるが、分母の $p(D)$（エビデンス）は尤度 $p(D \mid H)$ の周辺尤度（marginal likelihood）であり、周辺化（marginalization）計算をしなければならないので[8]、一般に算出が容易でない。ただしその

---

3　$m$ と $n$ が紛らわしいので注意。$m$ はデータの次元、すなわちデータ $x$ の持つカテゴリーの数である。$n$ はデータのサンプル数である。例えばカテゴリーが時間（期）しかなければ $m = 2$（時間と定数項）であるが、10分ごとの為替相場データを 250 日サンプルとしてデータ化すると $n = 36000$ となる。(15) 式の要素数は 36000 × 36000 という膨大な数になる。現実的に考えて $m$ あるいは (14) 式の $M$ はそう莫大な数になるとは考えられないが、データサンプル数はいくらでも大きくなりうるので、(15) 式で示されるカーネル行列の要素数は場合によっては計算不能なほど大きくなる可能性がある。

4　Shawe-Taylor 他 [5] 参照。カーネル法はガウス過程だけに使われるわけではない。

5　前掲 [10], p.36.

6　しかし脚注3 で示したように、それに代わって「サンプル数の呪い」が現れる。

7　確率変数が離散的な場合は確率、連続的な場合は確率密度と使い分けるべきであるが、煩雑さを避けるために以下ではすべて $p$ で表す。

8　例えば離散的な場合、$H_1, \cdots H_n$ が独立な仮定であるとすると $p(D \mid H_i)$ の周辺尤度 $p(D)$ は仮定 $H_i$ のすべてにわたる積和計算として

$$p(D) = p(D \mid H_i)p(H_i) + p(D \mid H_2)p(H_2) + \cdots + p(D \mid H_n)p(H_n)$$

のように、すなわち $p(H_i)$ による $p(D \mid H_i)$ の加重平均として求めなければならない。

計算が済んだものとすると、周辺尤度は$H$に関しては無関係な定数（$p(D)$）とみなすことができるので、(16) 式はしばしば

$$p(H \mid D) \propto p(D \mid H) \times p(H) \tag{17}$$

のように書かれる。

パラメトリック推定の考え方に従えばMAPとして求めるべき (17) 式の左辺は「仮定」であるパラメーター$\boldsymbol{w}$の事後分布であり

$$p(\boldsymbol{w} \mid D) = p(\boldsymbol{w} \mid \boldsymbol{X}, \boldsymbol{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}) \times p(\boldsymbol{w}) \tag{18}$$

と表される。

ベイズ的多重線形回帰モデルの代表的例として、(2) のモデルにさらに

$\varepsilon_i \sim N(0, \sigma^2)$ i.i.d（互いに独立に平均0分散$\sigma^2$の正規分布に従う）　(19)

の仮定と、$\boldsymbol{w}$の事前分布として$\boldsymbol{w} \sim N(\boldsymbol{w}_0, \boldsymbol{\Sigma})$：「平均$\boldsymbol{w}_0$共分散行列$\boldsymbol{\Sigma}$の多変量正規分布に従う」、すなわち

$$p(\boldsymbol{w}) = \frac{1}{\sqrt{(2\pi)^n |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_0)^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{w}_0)\right\} \tag{20}$$

という仮定を加える。また (2) と (19) より

$$\boldsymbol{y} \sim N(\boldsymbol{X}^T\boldsymbol{w}, \sigma^2\boldsymbol{I})$$

であるので、$\boldsymbol{y}$の尤度関数は

$$p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w}) = \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w}\right)\right\} \tag{21}$$

となり、(20) と (21) を (18) に代入すると$\boldsymbol{w}$の事後分布が次式のように計算できる。

$$\begin{aligned} p(\boldsymbol{w} \mid \boldsymbol{X}, \boldsymbol{y}) &\propto \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w}\right)^T\left(\boldsymbol{y}-\boldsymbol{X}^T\boldsymbol{w}\right)\right\} \\ &\cdot \exp\left\{-\frac{1}{2}(\boldsymbol{w}-\boldsymbol{w}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{w}-\boldsymbol{w}_0)\right\} \\ &= \exp\left[-\frac{1}{2}\left\{\frac{1}{\sigma^2}\left(\boldsymbol{w}^T\boldsymbol{T}\boldsymbol{T}^T\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{T}\boldsymbol{y}\right) + \boldsymbol{w}\boldsymbol{\Sigma}^{-1}\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{w}_0\right\}\right] \\ &\quad + \text{定数項} \\ &= \exp\left[-\frac{1}{2}\left\{\boldsymbol{w}^T\left(\frac{1}{\sigma^2}\boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{\Sigma}^{-1}\right)\boldsymbol{w} - 2\boldsymbol{w}^T\left(\frac{1}{\sigma^2}\boldsymbol{X}\boldsymbol{y} + \boldsymbol{\Sigma}^{-1}\boldsymbol{w}_0\right)\right\}\right] \\ &\quad + \text{定数項} \\ &= \exp\left[-\frac{1}{2}\left\{(\boldsymbol{w}-\bar{\boldsymbol{w}})^T\left(\frac{1}{\sigma^2}\boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{\Sigma}^{-1}\right)(\boldsymbol{w}-\bar{\boldsymbol{w}})\right\}\right] \\ &\quad + \text{定数項} \end{aligned} \tag{22}$$

ただし

$$\bar{\boldsymbol{w}} = \left(\frac{1}{\sigma^2}\boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\boldsymbol{X}\boldsymbol{y} + \boldsymbol{\Sigma}^{-1}\boldsymbol{w}_0\right) \tag{23}$$

と定義する。(22) の最後の行は2次式の平方完成を使う。(22) 式は明らかに$\boldsymbol{w} = \bar{\boldsymbol{w}}$の時最大となるから、$\boldsymbol{w}$の事後的MAP推定量は

$$\begin{aligned} p(\boldsymbol{w} \mid \boldsymbol{X}, \boldsymbol{y}) &= N\left(\bar{\boldsymbol{w}}, \left(\frac{1}{\sigma^2}\boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right) \\ &= N\left(\left(\frac{1}{\sigma^2}\boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\boldsymbol{X}\boldsymbol{y} + \boldsymbol{\Sigma}^{-1}\boldsymbol{w}_0\right), \right. \\ &\qquad \left. \left(\frac{1}{\sigma^2}\boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right) \end{aligned} \tag{24}$$

という多変量正規分布に従う。

回帰モデルが (2) のタイプではなく (7) のタイプの非線形モデルの場合は (24) 式の$\boldsymbol{X}$が$\boldsymbol{\Phi}$で置き換えられて

$$p(\boldsymbol{w} \mid \boldsymbol{X}, \boldsymbol{y}) = N\left(\left(\frac{1}{\sigma^2}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\frac{1}{\sigma^2}\boldsymbol{\Phi}\boldsymbol{y} + \boldsymbol{\Sigma}^{-1}\boldsymbol{w}_0\right), \left(\frac{1}{\sigma^2}\boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right) \tag{25}$$

となる。

### (iv)　パラメトリック予測の方法

予測とは既知のデザイン・データまたは訓練データの他に新たなインプット・データベクトル（テスト・ベクトル）$\boldsymbol{x}_*$が与えられた時にそれと対となる値すなわちテストの解答$\boldsymbol{y}_*$またはその分布$p(\boldsymbol{y}_*)$（ベイズ推定の場合）を求めることである。推定と予測は言葉として紛らわし

いが、推定は既知の訓練データと既知の観測値（教師ともいえる）を使って学習することであり（教師付き学習とも言う）、予測は学習の成果をもとに新たに出されたテストの解答を出すことに等しい。非ベイズ予測の場合、予測値はすでに計算された推定パラメーター (9) または (11) を使って

$$y_* = \boldsymbol{x}_*^T \hat{\boldsymbol{w}} \tag{26}$$

あるいは

$$y_* = \boldsymbol{\varphi}(\boldsymbol{x}_*)^T \hat{\boldsymbol{w}} \tag{27}$$

のように求められる。

ベイズ予測の場合求められるべきはそのような 1 点の値ではなく、すべての $\boldsymbol{w}$ の推測値を考慮した $y_*$ の「分布」ということになり、非線形回帰モデル (7) を例とすると、

$$\boldsymbol{S} = \left( \frac{1}{\sigma^2} \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Sigma}^{-1} \right) \tag{28}$$

および $D = (\boldsymbol{X}, \boldsymbol{y})$ と置いて、$\boldsymbol{w}$ の分布が

$$p(\boldsymbol{w} \mid D, \boldsymbol{x}_*) = N(\boldsymbol{w} \mid \bar{\boldsymbol{w}}, \boldsymbol{S}^{-1}) \tag{29}$$

の時、$y_*$ が次の予測モデルに従うならば

$$y_* = \boldsymbol{\varphi}(\boldsymbol{x}_*)^T \boldsymbol{w} + \eta \qquad \eta \sim N(0, \sigma_y^2) \tag{30}$$

すなわち $y_*$ の分布（尤度）が

$$p(y_* \mid \boldsymbol{w}, D, \boldsymbol{x}_*) = N(y_* \mid \boldsymbol{\varphi}(\boldsymbol{x}_*)^T \boldsymbol{w}, \sigma_y^2) \tag{31}$$

ならば、$p(y_* \mid D, \boldsymbol{x}_*) = p(y_* \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*)$ がどのような分布になるかということである。言い換えると $y_*$ の尤度 $p(y_* \mid \boldsymbol{w}, D, \boldsymbol{x}_*)$ の周辺尤度 $p(y_* \mid D, \boldsymbol{x}_*)$ を求めることであり

$$p(y_* \mid D, \boldsymbol{x}_*) = \int p(y_* \mid \boldsymbol{w}, D, \boldsymbol{x}_*)\, p(\boldsymbol{w} \mid D, \boldsymbol{x}_*) d\boldsymbol{w} \tag{32}$$

を計算することである。この式は予測分布（predictive distribution）とも呼ばれる。(32) 式は形式的にはあらゆる $\boldsymbol{w}$ の可能性を考慮し $p(\boldsymbol{w} \mid D, \boldsymbol{x}_*)$ を加重とする $p(y_* \mid \boldsymbol{w}, D, \boldsymbol{x}_*)$ の加重平均をとる（積分する）ということであるが、ここでは脚注で、Bishop[1], p.91以下で示されている代数的式変形を使って次のように求めることができる [9]。

$$\begin{aligned} p(y_* \mid D, \boldsymbol{x}_*) &= p(y_* \mid \boldsymbol{X}, \boldsymbol{y}, \boldsymbol{x}_*) \\ &= N\left( y_* \mid \boldsymbol{\varphi}(\boldsymbol{x}_*)^T \bar{\boldsymbol{w}}, \sigma_y^2 + \boldsymbol{\varphi}(\boldsymbol{x}_*)^T \boldsymbol{S}^{-1} \boldsymbol{\varphi}(\boldsymbol{x}_*) \right) \end{aligned} \tag{33}$$

ただし

$$\bar{\boldsymbol{w}} = \left( \frac{1}{\sigma^2} \boldsymbol{\Phi}\boldsymbol{\Phi}^T + \boldsymbol{\Sigma}^{-1}) \right)^{-1} \left( \frac{1}{\sigma^2} \boldsymbol{\Phi}\boldsymbol{y} + \boldsymbol{\Sigma}^{-1} \boldsymbol{w}_0 \right) \tag{34}$$

である。

---

9　$\boldsymbol{z} = (\boldsymbol{w}, y_*)^T$ と置くと、$p(\boldsymbol{z}) = p(\boldsymbol{w}, y_*)^T = p(y_* \mid \boldsymbol{w}) \cdot p(\boldsymbol{w})$, $\ln(p(\boldsymbol{z})) = \ln(p(y_* \mid \boldsymbol{w})) + \ln(p(\boldsymbol{w}))$, 定数部分は省略し exp の中の、$-(1/2)(\boldsymbol{w}-\bar{\boldsymbol{w}})^T (\boldsymbol{S}^{-1})^{-1} (\boldsymbol{w}-\bar{\boldsymbol{w}}) + (y_* - \boldsymbol{\varphi}^T \boldsymbol{w})^T \left(1 / \sigma_y^2\right) (y_* - \boldsymbol{\varphi}^T \boldsymbol{w})\}$ --(*) の 2 次部分だけを取り出すと

$$-\frac{1}{2} (\boldsymbol{w}, y_*) \begin{bmatrix} \boldsymbol{S} + \frac{1}{\sigma_y^2} \boldsymbol{\varphi}^T \boldsymbol{\varphi} & -\frac{\boldsymbol{\varphi}}{\sigma_y^2} \\ -\frac{\boldsymbol{\varphi}^T}{\sigma_y^2} & \frac{1}{\sigma_y^2} \end{bmatrix} \begin{bmatrix} \boldsymbol{w} \\ y_* \end{bmatrix}$$

Woodbury（Schur 補行列）公式:

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{C} & \boldsymbol{D} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{U} & -\boldsymbol{U}\boldsymbol{B}\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{U} & \boldsymbol{D}^{-1} + \boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{U}\boldsymbol{B}\boldsymbol{D}^{-1} \end{bmatrix} \quad \boldsymbol{U} = [\boldsymbol{A} - \boldsymbol{B}\boldsymbol{D}^{-1}\boldsymbol{C}]^{-1}$$

に当てはめると、上記 2 次部分の逆行列が $\boldsymbol{z}$ の共分散行列であるから。

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{S} + \frac{1}{\sigma_y^2} \boldsymbol{\varphi}^T \boldsymbol{\varphi} & -\frac{\boldsymbol{\varphi}}{\sigma_y^2} \\ -\frac{\boldsymbol{\varphi}^T}{\sigma_y^2} & \frac{1}{\sigma_y^2} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{S}^{-1} & \boldsymbol{S}^{-1}\boldsymbol{\varphi} \\ \boldsymbol{\varphi}^T \boldsymbol{S}^{-1} & \sigma_y^2 + \boldsymbol{\varphi}^T \boldsymbol{S}^{-1} \boldsymbol{\varphi} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{ww} & \boldsymbol{\Sigma}_{wy} \\ \boldsymbol{\Sigma}_{yw} & \boldsymbol{\Sigma}_{yy} \end{bmatrix}$$

すなわち $y_*$ の共分散行列は $\Sigma_{yy} = \sigma_y^2 + \boldsymbol{\varphi}^T \boldsymbol{S}^{-1} \boldsymbol{\varphi}$ である。一方 $\boldsymbol{z}$ の平均を $\mathrm{E}(\boldsymbol{z}) = (\mathrm{E}(\boldsymbol{w}), \mathrm{E}(y_*))^T$ と置いて $-(1/2)\{(\boldsymbol{z}\text{-}\mathrm{E}(\boldsymbol{z}))\, \Sigma^{-1}\, (\boldsymbol{z}-\mathrm{E}(\boldsymbol{z}))\}$ を展開すると $\boldsymbol{z}^T$ の一次の項は $\boldsymbol{z}^T \Sigma^{-1} \mathrm{E}(\boldsymbol{z})$ であり、(*) 式の一次の項は $(\boldsymbol{w}, y_*)^T \begin{bmatrix} \boldsymbol{S}\bar{\boldsymbol{w}} \\ 0 \end{bmatrix} = \boldsymbol{z}^T \begin{bmatrix} \boldsymbol{S}\bar{\boldsymbol{w}} \\ 0 \end{bmatrix}$ と書けるので、両者を比較すると $\Sigma^{-1}\, \mathrm{E}(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{S}\bar{\boldsymbol{w}} \\ 0 \end{bmatrix}$ すなわち $\mathrm{E}(\boldsymbol{z}) = \begin{bmatrix} \mathrm{E}(\boldsymbol{w}) \\ \mathrm{E}(y_*) \end{bmatrix} = \boldsymbol{\Sigma} \begin{bmatrix} \boldsymbol{S}\bar{\boldsymbol{w}} \\ 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{S}^{-1}\boldsymbol{S}\bar{\boldsymbol{w}} \\ \boldsymbol{\varphi}^T \boldsymbol{S}^{-1} \boldsymbol{S}\bar{\boldsymbol{w}} \end{bmatrix} = \begin{bmatrix} \bar{\boldsymbol{w}} \\ \boldsymbol{\varphi}^T \bar{\boldsymbol{w}} \end{bmatrix}$ 従って $\mathrm{E}(y_*) = \boldsymbol{\varphi}^T \bar{\boldsymbol{w}}$ となる。

非ベイズ予測における (26) (27) 式の $\hat{\boldsymbol{w}}$ もベイズ予測における (34) 式の $\bar{\boldsymbol{w}}$ も、その算出には訓練データと与えられた分散・共分散しか使われていないことが分かる。そこには過去のデータが変換されて圧縮されているが、新しく登場した新データ $\boldsymbol{x}_*$ の情報は反映されていない。予測にあたって、すでに推計されたパラメーターは新データによって更新されないままである。言い換えると学習はすでに完了しており、新データ（テストデータ）による改定はなされない。

### (v)　パラメトリック予測の問題点

以上が古典的なパラメトリック回帰モデルの方法であるが、よく知られている第1の問題点は過剰適合（over-fitting）の問題である。説明変数の数を増やしすぎると、関数が複雑化し滑らかさが失われるために、推計のパフォーマンスが落ちてしまうという問題である。この問題は時系列モデルの (4) から直感的に理解することができる。(4) でパラメーターを2つ $(w_1, w_2)$ にすると、推定関数は直線となり、データが直線という形で最も滑らかに平滑化されているが、$m$ を増やせば (4) 式はより波打つ $m-1$ 次曲線となる。しかし増やしすぎるとほぼデータの動きそのものに近くなり、データのランダム性を拾って将来が予測不可能となる。それを防ぐ方法として、例えば、最小二乗法の残差二乗和

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{w})^2$$

にペナルティ項として正則化項 $\lambda\boldsymbol{w}^T\boldsymbol{w}$ を付け加えて

$$\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{w})^2 + \lambda\boldsymbol{w}^T\boldsymbol{w}$$

とし、これを最小化するという「リッジ（Ridge）回帰」などがある[10]。リッジ回帰の場合 (9) は

$$\hat{\boldsymbol{w}} = \left(\boldsymbol{X}\boldsymbol{X}^T - \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}\boldsymbol{y}$$

(11) は

$$\hat{\boldsymbol{w}} = \left(\boldsymbol{\Phi}\boldsymbol{\Phi}^T - \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{\Phi}\boldsymbol{y}$$

と修正される。

パラメーター数の過剰を防止するためにAIC、BIC、wBIC[11] といった情報量基準を併用する方法もある。しかし、$\lambda$ を調整したり情報量基準に従えば必ず予測が精密になるかといえば、様々に言及されているように、ケース・バイ・ケースで、良し悪しは結果から判断するしかない。

推計の主目的が過去のデータへの全般的な当てはまりの良さや平滑性ではなく（over-fittingの回避はこの点に焦点が当てられている）、予測の精度であるとするならば、すなわち主目的は外挿であって内挿ではないとするならば、直近の仮想的予測の平均誤差率の少なさを基準としてパラメーター探索を行うという［10］のような発見的探索方法は実用的な観点からは、有用であると思われる。時間を説明変数とする多項式非線形回帰の例では、べきの次数を整数次数だけに限定せず、少数次数も含めて低次から連続的に増やしていくと[12]、大域的ではないが、局所的に、例えば2.35次といった次数で直近10期の平均誤差率が前後にくらべて極小になるといった、局所的最適点が発見できる。すなわちあらかじめパラメーターの個数を設定するのでなく、直近の平均誤差率が極小になるようにシミュレーションを通じて次数を細かく増やしていき、そうした中から大域的最小点を見つけるのである。モデル選択の cross validation（交差検証）に近い方法である。日経平均株価を例としたシミュレーションの結果、この方法で短期予測においてかなりの精度を持つ予測が実現されることが分かった［10］。

ただしこうした発見的探索方法は、局所的な極小が大域的な最小とは限らないという一般的

10 他に Lasso回帰、Elastic Net回帰などがあるが、本稿ではテーマからはずれるので、省略する。
11 Watanabe[7]
12 整数次のみで多項式回帰を行うと、極端な場合整数の次数が増えるごとに予測グラフの方向性が反転する。例えば2次の場合グラフの末端が上を向くが、3次になると下を向く。

な欠点を持っており、大域的最小に到達しないまま次数が増えていきオーバー・フローになってしまう可能性も高い。しかし本稿では省略するが、異なる局所的最適点を使って、波動の転換点（ゴールデンクロスやデッドクロス）の探知に応用することも可能である。実際低いべきの次数から大きい次数へと増やしていくということは、ローバスフィルターで周期の大きい波動を時系列から検知し、ブロックする周波数領域を変えながら次第に小周期の波動を検知していくという操作、例えて言うとラジオのチューニングと類似であり、さらなる応用の余地があると考えている。

問題の第2は、パラメトリック推計におけるパラメーターの「有限性」の問題である。滑らかな関数が理想であるにしても、平滑化に足りない部分を上例のように少数次のべきの項で細かく補完しても、有限個のパラメーターでは完全に近似できないかもしれない。実際、例えば3次+0.9999次のべきと 4次のべきの多項式回帰の間には埋めがたい溝がある。少数次のコンマ以下に 9 をいくら追加していっても 4次に滑らかに近づかない。前者のグラフは先端で上に上昇し、後者は下に下降するといった断絶があっても、少数次数の追加では溝を埋めることができない。

バラメーターの数が有限であるということは、特徴空間の中に写像した部分空間の次元が有限であるということに等しい。滑らかさを持ちかつ予測性に優れた真の関数は加算個のパラメーターどころか、非加算個のパラメーターまで拡張しなければ､求められないかもしれない。そうなると、無数のパラメーターの究極的な極限は実数値をとる連続な「関数」$f$そのものとなり、すなわち$f$が無限次元パラメーターとなり、(7) は

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i \qquad (35)$$

へと拡張される。推定すべき関数を表すのに、有限なパラメーター$\boldsymbol{w}$を持つ関数ではなく、実数値関数$f$にまで拡張する。これがガウス過程の考え方である。

問題の第3はすでに触れたが、パラメーターを算出する際のデータ利用の不完全性である。予測に必要となるパラメーターが新しいデータを使わず、過去のデータしか反映していないという、いわゆる「情報のボトルネック」である [13]。例えば (23) 式における $\bar{\boldsymbol{w}}$ のように、データによってパラメーターの事前分布が更新されるベイズ的予測においてすら、このボトルネックから免れることができない。

以上を要約すれば、パラメトリック予測の難点とは、結局既存のデータから計算された、更新されない有限の推定パラメーターを用いた関数で新たに与えられたデータに対する答えの予測を行うということに尽きる。例えて言うと、教科書を丸暗記したのに教科書に載ってない問題がテストに出たようなものである。以下ではガウス過程がこの軛をいかに振り払うかを示す。ただしそれでもハイパー・パラメーターと呼ばれるパラメーターからは逃れることができない、という結末を示す。

## Ⅱ　ガウス過程の概要

### (i)　ガウス過程の定義

確率測度が付与された標本空間の要素は一般的には事象 ($\omega$ とする) であり、例えば色とか出来事といった、数ではないものが含まれるが、ある条件 [14] のもとにそれらの事象$\omega$ に実数 (一般には実数ベクトル $\boldsymbol{x} \in R^n$) を対応させることができると、確率変数 (通常の変数$\boldsymbol{x}$ と区別してひとまず $\tilde{\boldsymbol{x}}$ と書くことにする) が得られる。その写像を確率写像$P(\omega)$ と呼ぶことにすると [15]、確率写像によって変数$\boldsymbol{x}$ に事象$\omega$ と結びついた確率的性質が与えられ、変数$\boldsymbol{x}$ は確定的な値を持つ数としてではなく確率変数$\tilde{\boldsymbol{x}}$ へと意味を変

13 Shawe-Taylor and Cristianini[5] (訳) , p.87.

14 $P : \Omega \to R^n$ を標本空間から $R^n$ への写像とするとき $R^n$ の $\sigma$加法族の要素 $Y$ から $\Omega$ への逆像 $P^{-1}(Y)$ が $\Omega$ の $\sigma$加法族の要素であるならば、この写像 $P(\Omega)$ の像は確率変数である。

える。事象$\omega$から実数への確率写像$P(\omega)$が定義されるのと同様に、関数空間$F$の要素$f$から確率的関数$\tilde{f}$への確率写像$\tilde{f}=P(f)$を定義することができる。確率的性質が加えられると、関数空間の要素としての関数$f$は確定的な値をとる関数ではなく、確率変数としての確率的関数$\tilde{f}$（「関数空間上の確率変数」とも呼ばれる。）へと変わる。すなわち関数自体が確率変数とみなされる。以下では煩雑さを避けるため確率的関数であっても$\tilde{f}$ではなく単に$f$と書く。また確率変数$\boldsymbol{x}$に対して確率分布が確率分布関数（例えば$G$とする）によって$G(\boldsymbol{x}),\ 0\le G(\boldsymbol{x})\le 1$のように定義されるのと同様に、確率的関数$f(\boldsymbol{x})$に対しても確率分布が確率分布関数によって$G(f(\boldsymbol{x})),\ 0\le G(f(\boldsymbol{x}))\le 1$のように定義される。これは$f(\boldsymbol{x})$を要素とする関数空間上の確率分布と呼ばれる。

$x_i(i=1,2\cdots,n)$が確率変数でそれぞれが正規分布（ガウス分布）に従うとき$n$個の要素からなるベクトル$\boldsymbol{x}=(x_1,x_2,\cdots,x_n)$は多変量正規分布に従う。このアナロジーとして

**「ガウス過程GPの定義」**

任意の有限な自然数$n$に対して$n$個の関数$(f(\boldsymbol{x}_1),f(\boldsymbol{x}_2),\cdots f(\boldsymbol{x}_n))$が多変量正規分布に従うとき、関数空間上の確率分布をガウス過程（Gaussian Process: GP）という[16]。

関数は$\boldsymbol{x}$上で定義されており、$\boldsymbol{x}$が異なればいくらでも無数の関数が存在する。関数を定義する点$\boldsymbol{x}$が$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots$のように増えていけば、それに応じて関数は$f(\boldsymbol{x}_1),f(\boldsymbol{x}_2),\cdots$と増えていくだけでなく、究極には非加算無限個すなわち実数直線的な濃度で隙間なく無限に増えていき、それに応じて$f(\boldsymbol{x})$の数は非加算無限個存在することになる。言い換えると関数空間の次元は本来無限である。上の定義は有限から無限への飛躍があるようにも見える。定義の前段が有限次元の点の上での$n$個の正規分布の集まりを条件としているが、後段ではそれを無限次元空間上の確率分布にまで拡張している。しかしながらこうした拡張は、「コルモゴロフの拡張定理」[17]により保障されており、他の確率過程の定義においても援用されるロジックである。すなわち無限次元上で成立しなければならない条件を言うのに、「任意の」有限次元の上で成立すればそれでよい、あるいは任意の有限次元で成立するならばその結果を無限次元に拡張できる、という考え方である。従ってガウス過程は有限点（つまりサンプル点）での有限次元多変量正規分布を条件にしつつも[18]、概念的には、究極的に非加算無限次元の関数空間上の多変量正規分布として定義できるのである。

正規分布は平均と共分散で決定されるので、ガウス過程の前段の条件は、$\boldsymbol{x}_i$の関数である平均関数

$$m(\boldsymbol{x}_i)=E(f(\boldsymbol{x}_i)) \tag{36}$$

と共分散関数

$$\begin{aligned} k(\boldsymbol{x}_i,\boldsymbol{x}_j)&=Cov\Big(f(\boldsymbol{x}_i),f(\boldsymbol{x}_j)\Big)\\ &=E\Big\{\Big(f(\boldsymbol{x}_i)-m(\boldsymbol{x}_i)\Big)\Big(f(\boldsymbol{x}_j)-m(\boldsymbol{x}_j)\Big)\Big\} \end{aligned} \tag{37}$$

を要素とする共分散行列

$$K\Big[\boldsymbol{x}_i,\boldsymbol{x}_j\Big]=\begin{bmatrix} k(\boldsymbol{x}_1,\boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1,\boldsymbol{x}_n)\\ \vdots & \vdots & \vdots\\ k(\boldsymbol{x}_n,\boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_n,\boldsymbol{x}_n)\end{bmatrix} \tag{38}$$

により、任意の有限な$n$に対して

---

15 Kolmogorov（邦訳）[9]では「確率関数」と名付けられている。ここではイメージのしやすい確率写像と呼ぶ。

16 ガウス過程の歴史は古く、Wiener[8]まで遡ると言われている。従ってその定義も必ずしも一律ではない。例えばRasmussen and Williams [4], p.13. ではガウス過程を分布ではなく、「確率変数の集まり」、ただしそのうちのどの有限個も結合ガウス分布を持つ、と定義している。この「確率変数」を「確率的関数」と読み替えれば、通常流布しているガウス過程の定義と一致する。

17 コルモゴロフ[9]第3章§4「無限次元空間の確率」

18 有限次元を条件とするのは、現実的にデータ（サンプル）は有限で、計算する際に現実に有限データしか存在しないからである、ともいえる。

$$\begin{bmatrix} f(\boldsymbol{x}_1) \\ \vdots \\ f(\boldsymbol{x}_n) \end{bmatrix} \sim N\left( \begin{bmatrix} m(\boldsymbol{x}_1) \\ \vdots \\ m(\boldsymbol{x}_n) \end{bmatrix}, \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \vdots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix} \right) \tag{39}$$

という条件が成り立つことを示している。ただし注意すべきは(36)(37)は定義上の仮定であって実際に漠然とした関数の期待値や共分散が初めから計算できるわけではない。飽くまでそういう「設定」である。従って事前の仮定であるならば、文字通りのデータからの共分散ではなく、共分散をエミュレートするようなカーネル関数$k(\boldsymbol{x}, \boldsymbol{x}')$を初期設定すればよい。すなわちガウス過程は仮定された期待値と、共分散と同等な性質（カーネル関数からなる行列が実対称行列となる）を生み出すカーネル関数を初期設定して運用される。従ってガウス過程の定義はガウス分布を$\mathcal{GP}$と表記して

$$f(\boldsymbol{x}) \sim \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')) \tag{40}$$

と表すこともできる。

ガウス過程の最大の特徴は、(39)式を見れば明らかなように推定目標としてのパラメーター$\boldsymbol{w}$とは無縁であるという点である。仮に推定すべきものがあるとすればそれは$f(\boldsymbol{x})$そのものであって、しかもこれから展開するように、予測において$f(\boldsymbol{x})$は(32)式と類似の「周辺化」によって姿を消す。ただし共分散関数(37)はカーネル関数へと一般化されるので、そのカーネル関数自体を具体的な関数として決めるためのパラメーターが必要となる。このハイパー・パラメーターは、初期設定としてその値を与えなければならない。その意味においてガウス過程は完全なパラメーター・フリーではない。

### (ii)　ガウス過程による予測

ガウス過程における回帰予測は次のような手順をとる。まずモデルとして

$$y = f(\boldsymbol{x}) + \varepsilon \tag{41}$$

という一般形の関数を考える。訓練データ$D = (\boldsymbol{X}, \boldsymbol{y})$が与えられると(41)は

$$y_i = f(\boldsymbol{x}_i) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_y^2), (i = 1, 2, \cdots, n) \tag{42}$$

のように離散的に表される。$f(\boldsymbol{x})$に関して事前に仮定されるものはカーネル関数$k(\boldsymbol{x}, \boldsymbol{x}')$から計算されるある種擬似的な共分散関数

$$\boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) = \begin{bmatrix} k(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_1, \boldsymbol{x}_n) \\ \vdots & \vdots & \vdots \\ k(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & k(\boldsymbol{x}_n, \boldsymbol{x}_n) \end{bmatrix} \tag{43}$$

と、$n$次元平均関数ベクトル

$$\boldsymbol{m}(\boldsymbol{x}) = \left(m(\boldsymbol{x}_1), m(\boldsymbol{x}_2), \cdots, m(\boldsymbol{x}_n)\right)^T = \boldsymbol{0} \tag{44}$$

である（(44)式の$\boldsymbol{0}$は事前の仮定に過ぎない）。(42)(43)(44)よりベクトル$\boldsymbol{f} = \left(f(\boldsymbol{x}_1), f(\boldsymbol{x}_2), \cdots, f(\boldsymbol{x}_n)\right)^T$の事前分布は

$$p(\boldsymbol{f} \mid \boldsymbol{X}) = N\left(\boldsymbol{f} \mid \boldsymbol{0}, \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X})\right) \tag{45}$$

である。これは(18)式右辺の$p(\boldsymbol{w})$に相当するが、パラメーター$\boldsymbol{w}$は関数$\boldsymbol{f}$に代わっている。

(42)より(18)式右辺の尤度$p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{w})$に相当するものは

$$p(\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{f}) = N(\boldsymbol{y} \mid \boldsymbol{f}, \sigma_y^2 \boldsymbol{I}) \tag{46}$$

である。

ガウス過程回帰による予測とは、訓練データ$D = (\boldsymbol{X}, \boldsymbol{y})$が分かっているとき新しいデータ（テスト・データ$\boldsymbol{X}_* = \left[\boldsymbol{x}_1^*, \boldsymbol{x}_2^*, \cdots, \boldsymbol{x}_h^*\right]$）が追加された場合それと対になる

$$\boldsymbol{y}_* = \left(y_1^*, y_2^*, \cdots, y_h^*\right)^T$$

の分布はどのようなものになるか、ということであり、

$$p(\boldsymbol{y}_* \mid \boldsymbol{X}, \boldsymbol{X}_*, \boldsymbol{y})$$

を求めることである。そのためまず既知のデータ$\boldsymbol{y}$の共分散を求めると、$f(\boldsymbol{x}_i)$と$\varepsilon_i$が無相関$\varepsilon_i$と$\varepsilon_i (i \neq j)$が無相関という仮定の下に(42)式から

$$\boldsymbol{K}(\boldsymbol{y}, \boldsymbol{y}) = \boldsymbol{K}(\boldsymbol{X}, \boldsymbol{X}) + \sigma_y^2 \boldsymbol{I} \tag{47}$$

が導かれる[19]。従ってガウス過程の定式によれ

ば $\boldsymbol{y}$ と $\boldsymbol{y}_*$ の同時分布が。

$$p(\mathbf{z}) = p\left(\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{y}_*\end{bmatrix}\right) = N\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{E}(\boldsymbol{y}_*)\end{bmatrix}, \begin{bmatrix}\boldsymbol{K}(\boldsymbol{y},\boldsymbol{y}) & \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*)\\ \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*)\end{bmatrix}\right)$$

$$= N\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{\mu}_*\end{bmatrix}, \begin{bmatrix}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\sigma_y^2\boldsymbol{I} & \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*)\\ \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}) & \boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*)\end{bmatrix}\right) \quad (48)$$

のように書かれる。$\boldsymbol{\mu}_*$ はデータ $\boldsymbol{y}$ の下での $\boldsymbol{y}_*$ の条件付き期待値 $\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}} = E(\boldsymbol{y}_* \mid \boldsymbol{y})$ ではないことに注意。事前的仮定として $\boldsymbol{\mu}_* = 0$ と置いても一般性を失わないが、$\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}}$ の一般式を求めるために $\boldsymbol{\mu}_*$ を (48) 式中に残しておく。ガウス過程の予測は、煩雑さを避けるため $\boldsymbol{X}$ を省けば、$\boldsymbol{y}$ の下で $\boldsymbol{y}_*$ の条件付き確率分布 $p(\boldsymbol{y}_* \mid \boldsymbol{y})$ を求めることであるとしても表現できる。以下この表現を使う。

予測式の導出を詳しく追ってみよう。なお簡単化のため

$$\begin{aligned}&\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\sigma_y^2\boldsymbol{I} = \boldsymbol{K}_1\\ &\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*) = \boldsymbol{K}_2\\ &\boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}) = \boldsymbol{K}_3\\ &\boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*) = \boldsymbol{K}_4\end{aligned}$$

と置く。また

$$\Sigma = \begin{bmatrix}\boldsymbol{K}_1 & \boldsymbol{K}_2\\ \boldsymbol{K}_3 & \boldsymbol{K}_4\end{bmatrix}$$

$$\Sigma^{-1} = \begin{bmatrix}\boldsymbol{K}_1 & \boldsymbol{K}_2\\ \boldsymbol{K}_3 & \boldsymbol{K}_4\end{bmatrix}^{-1} = \begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{bmatrix}$$

とする。

$$p(\boldsymbol{y}_* \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y},\boldsymbol{y}_*)}{p(\boldsymbol{y})} \quad (49)$$

の右辺、左辺のそれぞれの項を正規確率密度関数で表すと

$$p(\boldsymbol{y},\boldsymbol{y}_*) = c_1 \exp\left\{-\frac{1}{2}\left((\boldsymbol{y}^T,(\boldsymbol{y}_*-\boldsymbol{\mu}_*)^T)\Sigma^{-1}\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{y}_*-\boldsymbol{\mu}_*\end{bmatrix}\right)\right\}$$

$$p(\boldsymbol{y}) = c_2 \exp\left\{-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{X}_1^{-1}\boldsymbol{y}\right\}$$

$$p(\boldsymbol{y}_* \mid \boldsymbol{y}) = p(\boldsymbol{y},\boldsymbol{y}_*) \div p(\boldsymbol{y})$$

$$= c_3\left\{-\frac{1}{2}\left((\boldsymbol{y}^T,(\boldsymbol{y}_*-\boldsymbol{\mu}_*)^T)\Sigma^{-1}\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{y}_*-\boldsymbol{\mu}_*\end{bmatrix} - \boldsymbol{y}^T\boldsymbol{X}_1^{-1}\boldsymbol{y}\right)\right\}$$

$$= c_3\left\{-\frac{1}{2}\left((\boldsymbol{y}^T,(\boldsymbol{y}_*-\boldsymbol{\mu}_*)^T)\begin{bmatrix}\boldsymbol{A} & \boldsymbol{B}\\ \boldsymbol{C} & \boldsymbol{D}\end{bmatrix}\begin{bmatrix}\boldsymbol{y}\\ \boldsymbol{y}_*-\boldsymbol{\mu}_*\end{bmatrix} - \boldsymbol{y}^T\boldsymbol{X}_1^{-1}\boldsymbol{y}\right)\right\} \quad (50)$$

(50) 式の exp のかっこの中を $\boldsymbol{y}_*$ を含む項とそれ以外の項に分けると

$$= c_3 \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_*^T\boldsymbol{D}\boldsymbol{y}_* + \boldsymbol{y}_*^T(\boldsymbol{C}\boldsymbol{y}-\boldsymbol{D}\boldsymbol{\mu}_*) + (\boldsymbol{y}^T\boldsymbol{B}+\boldsymbol{\mu}_*^T)\boldsymbol{y}_*\right)\right.$$
$$\left. + (\boldsymbol{y}_*\text{を含まない項})\right\} \quad (51)$$

(51) 式の ($\boldsymbol{y}_*$ を含まない項) は $\boldsymbol{y}_*$ にとっては定数項となるので exp を考慮するとこれを先頭の定数に含めることができ、

$$p(\boldsymbol{y}_* \mid \boldsymbol{y}) = c_4 \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_*^T\boldsymbol{D}\boldsymbol{y}_* + \boldsymbol{y}_*^T(\boldsymbol{C}\boldsymbol{y}-\boldsymbol{D}\boldsymbol{\mu}_*)\right.\right.$$
$$\left.\left. + (\boldsymbol{y}^T\boldsymbol{B}+\boldsymbol{\mu}_*^T)\boldsymbol{y}_*\right)\right\} \quad (52)$$

となる。

一方条件付き共分散を $\boldsymbol{\Sigma}_{\boldsymbol{y}_*|\boldsymbol{y}}$ とすれば $G = \boldsymbol{\Sigma}_{\boldsymbol{y}_*|\boldsymbol{y}}^{-1}$ として $p(\boldsymbol{y}_* \mid \boldsymbol{y})$ は

$$p(\boldsymbol{y}_* \mid \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{y}_*-\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}})^T\boldsymbol{G}(\boldsymbol{y}_*-\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}})\right\}$$

$$= \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_*^T\boldsymbol{G}\boldsymbol{y}_* - \boldsymbol{y}_*^T\boldsymbol{G}\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}} - \boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}}^T\boldsymbol{G}\boldsymbol{y}_* + \boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}}^T\boldsymbol{G}\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}}\right)\right\} \quad (53)$$

とも書かれるはずである。(53) 式の最後の定数項 $\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}}^T\boldsymbol{G}\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}}$ は定数となって exp の前にでるから (53) は

$$p(\boldsymbol{y}_* \mid \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}\left(\boldsymbol{y}_*^T\boldsymbol{G}\boldsymbol{y}_* - \boldsymbol{y}_*^T\boldsymbol{G}\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}} - \boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}}^T\boldsymbol{G}\boldsymbol{y}_*\right)\right\} \quad (54)$$

19 ノイズの相関性がある場合 (47) 式はノイズの共分散 $\boldsymbol{\Sigma}$ を用いて $\boldsymbol{K}(\boldsymbol{y},\boldsymbol{y}) = \boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}) + \boldsymbol{\Sigma}$ のように書くことができる。Rasmussen 他 [4], p.190, 9.2 Noise Modes with Dependencies を参照。

ということになる。(52) と (54) の $\boldsymbol{y}_*$ の 2 次項を比較すると

$$\boldsymbol{G}=\boldsymbol{D} \tag{55}$$

であり $\boldsymbol{y}_*^T$ の係数を比較すると

$$-\boldsymbol{G}\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}}=(\boldsymbol{C}\boldsymbol{y}-\boldsymbol{D}\boldsymbol{\mu}_*)$$

$|\boldsymbol{G}|\neq 0$ 従って (55) より $|\boldsymbol{D}|\neq 0$ を仮定すると

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}}&=-\boldsymbol{G}^{-1}(\boldsymbol{C}\boldsymbol{y}-\boldsymbol{D}\boldsymbol{\mu}_*)=-\boldsymbol{G}^{-1}\boldsymbol{C}\boldsymbol{y}+\boldsymbol{G}^{-1}D\boldsymbol{\mu}_*\\&=-\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{y}+\boldsymbol{D}^{-1}\boldsymbol{D}\boldsymbol{\mu}_*=\boldsymbol{\mu}_*-\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{y}\end{aligned} \tag{56}$$

てある。

ここで Woodbury 公式

$$\begin{aligned}&\boldsymbol{A}=\boldsymbol{K}_1^{-1}+\boldsymbol{K}_1^{-1}\boldsymbol{K}_2\mathbf{V}^{-1}\boldsymbol{K}_3\boldsymbol{K}_1^{-1}\\&\boldsymbol{B}=-\boldsymbol{K}_1^{-1}\boldsymbol{K}_2\mathbf{V}^{-1}\\&\boldsymbol{C}=-\mathbf{V}^{-1}\boldsymbol{K}_3\boldsymbol{K}_1^{-1}\\&\boldsymbol{D}=\boldsymbol{V}^{-1}\\&\boldsymbol{V}=\left[\boldsymbol{K}_4-\boldsymbol{K}_3\boldsymbol{K}_1^{-1}\boldsymbol{K}_2\right]\end{aligned}$$

を使うと

$$\boldsymbol{\Sigma}_{\boldsymbol{y}_*|\boldsymbol{y}}=\boldsymbol{G}^{-1}=\boldsymbol{D}^{-1}=\boldsymbol{V}=\boldsymbol{K}_4-\boldsymbol{K}_3\boldsymbol{K}_1^{-1}\boldsymbol{K}_2 \tag{57}$$

および

$$\begin{aligned}\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}}&=\boldsymbol{\mu}_*-\boldsymbol{D}^{-1}\boldsymbol{C}\boldsymbol{y}=\boldsymbol{\mu}_*-\boldsymbol{V}\left(-\mathbf{V}^{-1}\boldsymbol{K}_3\boldsymbol{K}_1^{-1}\right)\boldsymbol{y}\\&=\boldsymbol{\mu}_*+\boldsymbol{K}_3\boldsymbol{K}_1^{-1}\boldsymbol{y}\end{aligned} \tag{58}$$

となる。従って $\boldsymbol{y}_*$ の条件付き分布は

$$p(\boldsymbol{y}_*|\boldsymbol{y})=N(\boldsymbol{y}_*\mid\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}},\Sigma_{\boldsymbol{y}_*|\boldsymbol{y}}) \tag{59}$$

$$\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}}=\boldsymbol{\mu}_*+\boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X})\left[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\sigma_y^2\boldsymbol{I}\right]^{-1}\boldsymbol{y} \tag{60}$$

$$\begin{aligned}\Sigma_{\boldsymbol{y}_*|\boldsymbol{y}}&=\boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X}_*)-\boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X})\\&\cdot\left[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\sigma_y^2\boldsymbol{I}\right]^{-1}\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X}_*)\end{aligned} \tag{61}$$

となり、その期待値も共分散も既存のデータだけでなく、更新されたデータ $\boldsymbol{X}_*$ を含むことが分かる。ガウス過程においては新たに付け加えられたテスト・データの情報は予測に際して完全に利用され、情報のボトルネックの問題は生じないのである。

初期設定として $\boldsymbol{\mu}_*=0$ を仮定すると (60) 式は

$$\boldsymbol{\mu}_{\boldsymbol{y}_*|\boldsymbol{y}}=\boldsymbol{K}(\boldsymbol{X}_*,\boldsymbol{X})\left[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\sigma_y^2\boldsymbol{I}\right]^{-1}\boldsymbol{y} \tag{62}$$

となる。

なお予測のためのテスト・データが $\boldsymbol{X}_*=\left[\boldsymbol{x}_1^*,\boldsymbol{x}_2^*,\cdots,\boldsymbol{x}_h^*\right]$ のような複数点でなく、一点 $\boldsymbol{x}_*$ の場合、求めるものはスカラー $y_*$ の条件付き分布 $p(y_*\mid\boldsymbol{y})$ であり、(60) は

$$\boldsymbol{k}_*=\begin{bmatrix}k(\boldsymbol{x}_*,\boldsymbol{x}_1)\\k(\boldsymbol{x}_*,\boldsymbol{x}_2)\\\vdots\\k(\boldsymbol{x}_*,\boldsymbol{x}_n)\end{bmatrix}$$

と置いてスカラー値

$$\mu_{\boldsymbol{y}_*|\boldsymbol{y}}=\mu_*+\boldsymbol{k}_*^T\left[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\sigma_y^2\boldsymbol{I}\right]^{-1}\boldsymbol{y} \tag{63}$$

初期設定を $\mu_*=0$ と仮定する場合、(63) はスカラー値

$$\mu_{\boldsymbol{y}_*|\boldsymbol{y}}=\boldsymbol{k}_*^T\left[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\sigma_y^2\boldsymbol{I}\right]^{-1}\boldsymbol{y} \tag{64}$$

となる。また (61) はスカラー値

$$Var(\boldsymbol{x}_*)=k(\boldsymbol{x}_*,\boldsymbol{x}_*)-\boldsymbol{k}_*^T\left[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X})+\sigma_y^2\boldsymbol{I}\right]^{-1}\boldsymbol{k}_* \tag{65}$$

となる。

この場合 1 点 (の分布) を予測するので次はどの点を予測するのが最適かという exploit (活用) と explore (探索) の間の最適配分の問題が新たに生じることになる [20]。

### (iii)　ハイパー・パラメーター

ノンパラメトリック予測は以上のようにパラメーターそのものを推定するのではないという意味においてはノンパラメトリックではあるが、関数の分布の共分散行列 (43) あるいは共分散関数 (カーネル関数 $\boldsymbol{k}(\boldsymbol{x},\boldsymbol{x}')$ ) が初期設定として仮定されなければならない、という意味

20 Lévesque, Julien-Charles 他 [3], p.9.

においてはパラメーターから完全に自由ではない。なぜならば初期設定といえども関数を特定するためにはパラメーターが必要だからである。例えば有名な動径基底関数（RBF: Radial Basis Function）は

$$k(\boldsymbol{x},\boldsymbol{x}')=\sigma^2\exp\left(\frac{(\boldsymbol{x}-\boldsymbol{x}')^T(\boldsymbol{x}-\boldsymbol{x}')}{2\lambda^2}\right)$$

のように定義されるが、明らかに2つのパラメーター$\sigma$と$\lambda$を持ち、これらを定めない限り不定になる。こうした初期設定としてのパラメーターはすでに触れたようにハイパー・パラメーター（Hyper Parameter）と呼ばれる。ハイパー・パラメーターが最適に決められないと、結局はover-fittingやその逆の過剰な平滑化になり、予測の失敗に繋がるかもしれない。

共分散行列（カーネル）(43)の条件は、行列が対称あるいは半正定値であるということだけであり、その条件を満たせば様々なカーネル関数の設計が可能である。また訓練データの波動の特徴に応じて、複数のカーネル関数の足し算掛け算によっていくらでも新しいカーネル関数を創造することができる[21]。するとモデルには様々なハイパー・パラメーター$\boldsymbol{\theta}$が混在することになる。しかしそれらの値をどのように与えたらよいであろうか。事前の仮定にすぎない値（例えば任意の定数とする）も実際のデータが与えられれば、事後的に修正されるであろうが、その結果は思いもよらない予測値をもたらすかもしれない。そこで訓練データ$D=(\boldsymbol{X},\boldsymbol{y})$の下で出現する$\boldsymbol{\theta}$の事後確率$p(\boldsymbol{\theta}\mid D)=p(\boldsymbol{\theta}\mid\boldsymbol{X},\boldsymbol{y})$を最大にするような$\boldsymbol{\theta}$の値を選ぶのがよいであろう。すなわちベイズ定理に従うと

$$p(\boldsymbol{\theta}|\boldsymbol{X},\boldsymbol{y})=\frac{p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})\,p(\boldsymbol{\theta}|\boldsymbol{X})}{p(\boldsymbol{y}|\boldsymbol{X})}\qquad(66)$$

を最大にしたい。そのためには右辺の分子の尤度$p(\boldsymbol{y}\mid\boldsymbol{X},\boldsymbol{\theta})$を最大にすればよい。これがガウス過程における「タイプ2の尤度最大化問題」である。すなわち$\boldsymbol{\theta}$を変数として

$$\begin{aligned}L(\boldsymbol{\theta})&=-ln\left(p(\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\theta})\right)\\&=\frac{1}{2}\boldsymbol{y}^T\left[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X},\boldsymbol{\theta})+\sigma_y^2\boldsymbol{I}\right]^{-1}\boldsymbol{y}\\&\quad+\frac{1}{2}ln\left(\det\left(\left[\boldsymbol{K}(\boldsymbol{X},\boldsymbol{X},\boldsymbol{\theta})+\sigma_y^2\boldsymbol{I}\right]\right)\right)+C\end{aligned}\qquad(67)$$

を最大化する問題ということになる。(67)式は一般に複数の局所解を持ち、また逆行列の計算量が膨大になるため、最適化のためのいくつかの数値計算法が提案されている。また代替的な方法としてシミュレーション的な予測を繰り返すことによってハイパー・パラメーターの初期値を模索するという交差確認法cross validationの手法も可能である。

以上を要約するとガウス過程を実際のデータに対して実行するには

(1) どのようなカーネル関数をどのように組み合わせで使うか（モデル選択）
(2) いかにして最適なハイパー・パラメーターを求めるか[22]

という2つの作業が必要となってくる。問題はいわば先延ばしにされたような形である。これらの問題及び、時系列データを使ったパラメーター予測とノンパラメーター予測（ガウス過程）のパフォーマンス比較は、次稿以降で行う。

**［参考文献］**

[1] Bishop, Christopher M., *Pattern Recognition and Machine Learning*, Springer, 2006（『パターン認識と機械学習（ベイズ理論による統計的予測）　上下 元田浩ほか監訳　丸善出版　2012』）.

[2] Duvenaud, David Kristjanson, *Automatic Model Construction with Gaussian Processes* (dissertation for Doctor of Philosophy,

21 Duvenaud[2]参照。モデル選択（Model Selection）と呼ばれている。
22 Lévesque, Julien-Charles[3] 参照。

University of Cambridge), https://www.cs.toronto.edu/~duvenaud/thesis.pdf, 2014.

[3] Lévesque, Julien-Charles, *Bayesian Hyperparameter Optimization: Overfitting, Ensembles and Conditional Spaces* (Ph. D. thesis), https://www.etsmtl.ca/getattachment/Unites-de-recherche/LIVIA/Recherche-et-innovation/Theses/JC-Levesque_PhD_2018.pdf, 2018.

[4] Rasmussen, Carl Edward and Williams, Christopher K. I., *Gaussian Processes for Machine Learning*, The MIT Press, 2006.

[5] Shawe-Taylor, John and Cristianini, Nello, *Kernel methods for Pattern analysis*, Cambridge University Press, 2004 (『カーネル法によるパターン解析』大北剛訳　共立出版　2010).

[6] Theodoridis, Sergios, *Machine Learning—A Bayesian and Optimization Perspective*, Elsevier, 2015.

[7] Watanabe, Sumio, 'A Widely Applicable Bayesian Information Criterion', *Journal of Machine Learning Research*, 14, 2013, pp.867-897.

[8] Wiener, N., *Extrapolation, Interpolation and Smoothing of Stationary Time Series*, MIT Press, 1949.

[9] コルモゴロフ，A. H.『確率論の基礎概念（第2版）』根本伸司　訳 東京図書1975.

[10] 荒木勝啓「日経平均への最適化非線形予測の適用」駒澤大学経済学論集 第47巻 第4号2016, pp.35-48.