6.4 ガウス過程

PRML 輪読会

髙橋 優輝

2023.11.15

大阪大学大学院情報科学研究科

目次

導入

定義-確率過程-

まず，確率過程とガウス過程の定義を確認する．ガウス過程については，この解説[1]を参考にした．

確率過程

$x \in \mathcal{X}$ に対して，$y(x)$ を確率変数とする．このとき，次の集合を確率過程という．

$$\{y(x) \mid x \in \mathcal{X}\}$$

[1]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 ガウス過程回帰の基礎 赤穂昭太郎

定義-ガウス過程-

ガウス過程

$\forall i\,(x_i \in \mathcal{X}), m\colon \mathcal{X} \to \mathbb{R}, k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ として，さらに，次式を定義する．

$$
\begin{aligned}
\boldsymbol{x}_N &\coloneqq (x_1, \ldots, x_N) \\
\boldsymbol{y}(\boldsymbol{x}_N) &\coloneqq (y(x_1), \ldots, y(x_N))^\top, \\
\boldsymbol{m}(\boldsymbol{x}_N) &\coloneqq (m(x_1), \ldots, m(x_N))^\top, \\
\boldsymbol{V}(\boldsymbol{x}_N, \boldsymbol{x}_N) &\coloneqq \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_N) \\ \vdots & \ddots & \vdots \\ k(x_N, x_1) & \cdots & k(x_N, x_N) \end{bmatrix}
\end{aligned}
$$

このとき，平均関数 m, 共分散関数 k によって定義される確率過程で，任意の n 点 $\boldsymbol{x}_N$ の関数値 $\boldsymbol{y}(\boldsymbol{x}_N)$ の分布が平均 $\boldsymbol{m}(\boldsymbol{x}_N)$，分散共分散行列 $V(\boldsymbol{x}_N, \boldsymbol{x}_N)$ の多次元ガウス分布に従うものをガウス過程という．（ただし，k はカーネル関数である．）

ガウス過程 is everywhere...

ガウス過程と等価なモデルはさまざまな分野で研究されている．

地球統計学においては，ガウス過程による回帰はクリギングとして知られている．

自己回帰移動平均モデル，カルマンフィルタ，RBF ネットワークなどもガウス過程の一種として見ることができる．

6.4.1 線形回帰再訪

線形回帰再訪-線形回帰の例での予測分布の再導出 (1/2)-

ここでは，線形回帰モデルがガウス過程の一例となっていることを確認する．

$$y(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}) \qquad (\boldsymbol{w} \in \mathbb{R}^M, \phi_j(\boldsymbol{x}) \text{: 基底関数}) \tag{6.49}$$

$\boldsymbol{w}$ の事前分布は等方的なガウス分布であるとする．(α: 分布の精度を表す HP)

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{0}, \alpha^{-1}\boldsymbol{I}) \tag{6.50}$$

また，訓練データの点集合 $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ における関数の値を要素に持つベクトル $\boldsymbol{y}$ を定義する．($\Phi_{ij} = \phi_j(\boldsymbol{x}_i)$)

$$\boldsymbol{y} \coloneqq \begin{bmatrix} y(\boldsymbol{x}_1) & \cdots & y(\boldsymbol{x}_N) \end{bmatrix}^\top = \boldsymbol{\Phi}\boldsymbol{w} \tag{6.51}$$

このとき，$\boldsymbol{y}$ の従う確率分布について考える．まず，式 (6.50) と式 (6.51) より，$\boldsymbol{y}$ はガウス分布に従う．

線形回帰再訪-線形回帰の例での予測分布の再導出 (2/2)-

$\boldsymbol{y}$ の平均と共分散は，それぞれ，次式で与えられる．

$$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{\Phi}\mathbb{E}[\boldsymbol{w}] = \boldsymbol{0} \tag{6.52}$$

$$\mathrm{cov}[\boldsymbol{y}] = \mathbb{E}[\boldsymbol{y}\boldsymbol{y}^\top] = \boldsymbol{\Phi}\mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top]\boldsymbol{\Phi}^\top = \alpha^{-1}\boldsymbol{\Phi}\boldsymbol{\Phi}^\top =: \boldsymbol{K} \tag{6.53}$$

ここで，$\boldsymbol{K} \in \mathbb{R}^{N\times N}$ の ij 要素は以下で与えられる．

$$K_{ij} = \alpha^{-1}\phi(\boldsymbol{x}_i)^\top\phi(\boldsymbol{x}_j) \tag{6.54}$$

よって，$\boldsymbol{y} \sim \mathcal{N}(\boldsymbol{0}, K)$ であり，線形回帰モデルはガウス過程の一例となっている．

線形回帰再訪-カーネル関数の設計 (1/2)-

ほとんどの応用について，$y(\boldsymbol{x})$ の平均についての事前知識はないため，対称性からこれを 0 とすることが多い（式 (6.52) より，これは，$\boldsymbol{w}$ の事前分布の平均を 0 とすることに対応）．このとき，ガウス過程は次のカーネル関数（共分散関数）のみで定義される．

$$\mathbb{E}[y(\boldsymbol{x}_i)y(\boldsymbol{x}_j)] = k(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

ここで，$\boldsymbol{w}$ の事前分布として，式 (6.50) を満たす線形回帰モデル (6.49) によって定義されるガウス過程の場合，カーネル関数は，次のようになる．

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) \coloneqq \alpha^{-1}\boldsymbol{\phi}(\boldsymbol{x}_i)^\top \boldsymbol{\phi}(\boldsymbol{x}_j)$$

カーネル関数はこのように基底関数から求めることも可能であるが，直接定義することも可能である．

$\mathbb{E}[\boldsymbol{y}] = \boldsymbol{\Phi}\mathbb{E}[\boldsymbol{w}] = \boldsymbol{0}$ (6.52)，　$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{0}, \alpha^{-1}\boldsymbol{I})$ (6.50)，　$y(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x})$ (6.49)

線形回帰再訪-カーネル関数の設計 (2/2)-

カーネル関数を定義することで，直接的に $\mathrm{cov}(\boldsymbol{y})$ を定義することができる．

図 6.4 は 2 種類のカーネル関数（ガウスカーネル，指数カーネル）を使用したときに，ガウス過程からサンプルされた関数を表している．

$$k_{\text{ガウス}}(\boldsymbol{x}, \boldsymbol{x}') := \exp(-\|\boldsymbol{x} - \boldsymbol{x}'\|^2/2\sigma^2) \tag{6.23}$$

$$k_{\text{指数}}(\boldsymbol{x}, \boldsymbol{x}') := \exp(-\theta\|\boldsymbol{x} - \boldsymbol{x}'\|) \tag{6.56}$$

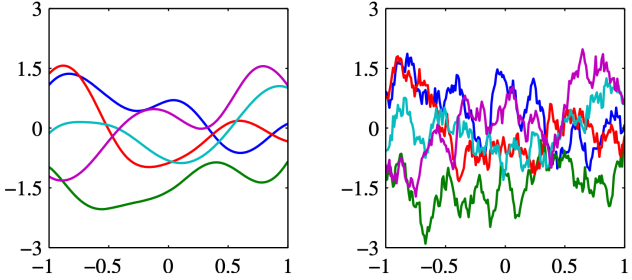


図 6.4 左：ガウスカーネル，右：指数カーネル 色の対応は何を意味している？

6.4.2 ガウス過程による回帰

ガウス過程による回帰-仮定-

ガウス過程を回帰問題に適用するには，次のように観測される目標変数の値に含まれるノイズを考える必要がある．

$$t_n = y_n + \epsilon_n \tag{6.57}$$

ここで，$y_n \coloneqq y(\boldsymbol{x}_n)$ であり，ϵ_n は n 番目の観測値に加えられるノイズで，それぞれの観測値に対して独立に決定されるとする．すなわち，次式が成り立つとする．

$$p(t_n \mid y_n) = \mathcal{N}(t_n \mid y_n, \beta^{-1}) \tag{6.58}$$

ここで，β はノイズの精度を表す HP である．

ガウス過程による回帰-周辺分布の導出-

ノイズの独立性より，$\boldsymbol{y} \coloneqq \left[y_1, \ldots, y_N\right]^\top, \boldsymbol{t} \coloneqq \left[t_1, \ldots, t_N\right]^\top$ とすると，次式が成立する．

$$p(\boldsymbol{t} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{t} \mid \boldsymbol{y}, \beta^{-1}\boldsymbol{I}) \tag{6.59}$$

また，周辺分布 $p(\boldsymbol{y})$ は次式で与えられる．（$\boldsymbol{K}$：グラム行列）

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{0}, \boldsymbol{K}) \tag{6.60}$$

このとき，式 (2.115)[2] を用いると，周辺分布 $p(\boldsymbol{t})$ は次式で与えられる．

$$p(\boldsymbol{t}) = \int p(\boldsymbol{t} \mid \boldsymbol{y}) p(\boldsymbol{y}) d\boldsymbol{y} = \mathcal{N}(\boldsymbol{t} \mid \boldsymbol{0}, \boldsymbol{C}) \tag{6.61}$$

ここで，共分散行列 $\boldsymbol{C}$ の ij 成分は次式で与えられる．

$$C_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \beta^{-1}\delta_{ij} \tag{6.62}$$

[2] $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}) \implies p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^\top)$

ガウス過程による回帰-ガウス過程回帰に用いるカーネル関数-(1/2)

ガウス過程回帰に用いるカーネル関数として次式が広く用いられている．

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \theta_0 \exp\left\{-\frac{\theta_1}{2}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2\right\} + \theta_2 + \theta_3 \boldsymbol{x}_i^\top \boldsymbol{x}_j \tag{6.63}$$

図 6.5 はこの共分散関数を式 (6.63) で定義して，パラメータ $\theta_0, \ldots, \theta_3$ の値を変えたときにサンプルされた関数である．

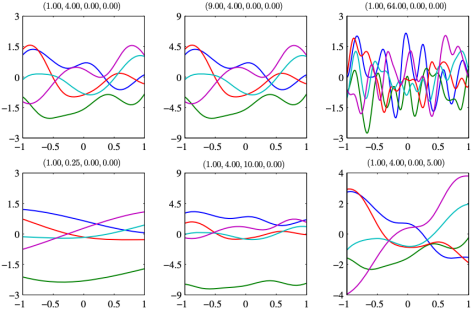


図 6.5 グラフの上：$(\theta_0, \theta_1, \theta_2, \theta_3)$

ガウス過程による回帰-ガウス過程回帰に用いるカーネル関数-(2/2)

図 6.6 は同時分布 (6.60) からのサンプルを，対応する値 (6.61) と共に示したものである．

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{0}, \boldsymbol{K}) \tag{6.60}$$

$$p(\boldsymbol{t}) = \mathcal{N}(\boldsymbol{t} \mid \boldsymbol{0}, \boldsymbol{C}) \tag{6.61}$$

$$C_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \beta^{-1}\delta_{ij} \tag{6.62}$$

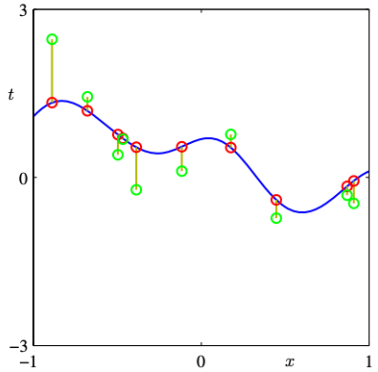


図 6.6
青線：ガウス過程による事前分布からサンプルされた関数
赤丸：入力集合 $\{\boldsymbol{x}_n\}$ に対応する値 y_n
緑丸：y_n にガウスノイズが加わった値 t_n

ガウス過程による回帰-新しい入力ベクトルに対する目標変数の予測 (1/3)-

回帰の目的は，訓練データの集合 $\{(\boldsymbol{x}_1, t_1), \ldots, (\boldsymbol{x}_N, t_N)\}$ が与えられた時に，新しい入力ベクトル $\boldsymbol{x}_{N+1}$ に対応する目標変数 t_{N+1} を予測することである．そのためには，予測分布 $p(t_{N+1} \mid \boldsymbol{t}_N)$ を求める必要がある．ここで，この予測分布は $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ と $\boldsymbol{x}_{N+1}$ に依存するが，表記を単純化するため，これらを省くことにする．

予測分布 $p(t_{N+1} \mid \boldsymbol{t}_N)$ を求めるために，同時分布 $p(\boldsymbol{t}_{N+1})$ を求める．ここで，$\boldsymbol{t}_{N+1} \coloneqq \left[t_1, \ldots, t_N, t_{N+1}\right]^\top$ である．

ガウス過程による回帰-新しい入力ベクトルに対する目標変数の予測 (2/3)-

式 (6.61),(6.62)[3] から，同時分布は次式で与えられる．

$$p(\boldsymbol{t}_{N+1}) = \mathcal{N}(\boldsymbol{t}_{N+1} \mid \boldsymbol{0}, \boldsymbol{C}_{N+1}) \tag{6.64}$$

$$\boldsymbol{C}_{N+1} \coloneqq \begin{bmatrix} \boldsymbol{C}_N & \boldsymbol{k} \\ \boldsymbol{k}^\top & c \end{bmatrix} \tag{6.65}$$

$$\boldsymbol{k} \coloneqq \begin{bmatrix} k(\boldsymbol{x}_1 & \boldsymbol{x}_{N+1}) & \cdots & k(\boldsymbol{x}_N, \boldsymbol{x}_{N+1}) \end{bmatrix}^\top$$

$$c \coloneqq k(\boldsymbol{x}_{N+1}, \boldsymbol{x}_{N+1}) + \beta^{-1}$$

これと $p(\boldsymbol{t}) = \mathcal{N}(\boldsymbol{t} \mid \boldsymbol{0}, \boldsymbol{C})$ (6.61) より，式 (2.81),(2.82) の結果を用いると，条件付き分布 $p(t_{N+1} \mid \boldsymbol{t})$ はガウス分布となり，その平均と共分散は次式で与えられることが分かる．

$$m(\boldsymbol{x}_{N+1}) = \boldsymbol{k}^\top \boldsymbol{C}_N^{-1} \boldsymbol{t} \tag{6.66}$$

$$\sigma^2(\boldsymbol{x}_{N+1}) = c - \boldsymbol{k}^\top \boldsymbol{C}_N^{-1} \boldsymbol{k} \tag{6.67}$$

[3] $p(\boldsymbol{t}) = \mathcal{N}(\boldsymbol{t} \mid \boldsymbol{0}, \boldsymbol{C})$ (6.61), $C_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \beta^{-1}\delta_{ij}$ (6.62)

ガウス過程による回帰-新しい入力ベクトルに対する目標変数の予測 (3/3)-

$\boldsymbol{k}$ と c はテスト点の入力 $\boldsymbol{x}_{N+1}$ に依存するため，予測分布の平均と分散は $\boldsymbol{x}_{N+1}$ に依存することが分かる．

図 6.7 は訓練データとテストデータが 1 つずつの場合のガウス過程による回帰の仕組みを表している．

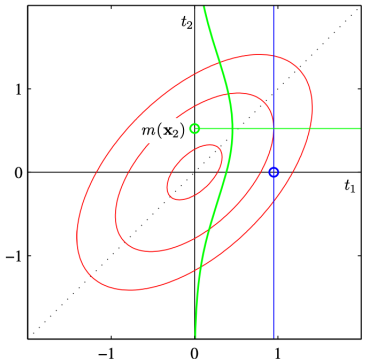


図 6.7 赤楕円：同時分布 $p(t_1, t_2)$ の等高線，青丸：訓練データ点 t_1，緑曲線：予測分布 $p(t_2 \mid t_1)$

ガウス過程による回帰-ガウス過程による回帰の例-

図 6.8 はガウス過程による回帰の例を示している．

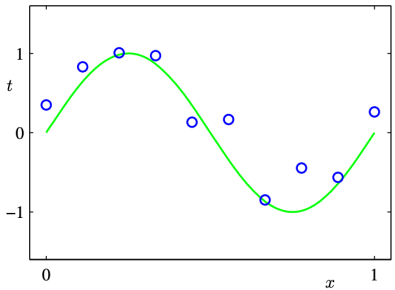


図 A.6
緑線：正弦関数
青丸：正弦関数にガウス分布に従うノイズを加えてサンプリングされた点

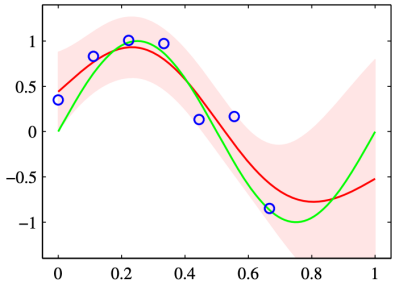


図 6.8 図 A.6 から一番右の 3 つのデータを取り除いたものに対して，ガウス過程を適用した結果
赤線：予測分布の平均
赤薄領域：赤線から標準偏差の 2 倍までの領域

ガウス過程による回帰-カーネル関数の制約-

カーネル関数についての唯一の制約は，式 (6.62)[4] で与えられる共分散行列が正定値でなければならないことである．λ_i をグラム行列 $\boldsymbol{K}$ の固有値とすると，$\boldsymbol{C}$ の対応する固有値は，$\lambda_i + \beta_i$ となる．$\beta > 0$ より，$\lambda_i \geq 0 \, \forall i \iff \boldsymbol{K} \succeq \boldsymbol{O} \implies \boldsymbol{C} \succ \boldsymbol{O}$ が分かる．よって，この制約は 6.2 節で議論したものと同じであるので，$\boldsymbol{K}$（すなわちカーネル関数 k）は 6.2 節で用いたテクニックを用いて設計すれば良い．

[4] $C_{ij} = k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \beta^{-1}\delta_{ij} \quad (6.62)$

ガウス過程による回帰-得手不得手-

ガウス過程を実際に計算する上で最も大きな計算量を要する部分が，行列 $\boldsymbol{C}_N$ の逆行列の計算であり，$O(N^3)$ の計算量がかかる．一方，基底関数を用いたモデルでは，行列 $\boldsymbol{S}_N$ の逆行列の計算がボトルネックで $O(M^3)$ の計算量が必要である．どちらにおいても，逆行列の計算は与えられた訓練集合に対して 1 回行う必要があり，新しいテスト点が与えられると，行列ベクトル積の計算が必要で，ガウス過程では $O(N^2)$，線形の基底関数モデルでは $O(M^2)$ で与えられる共分散行列が正定値でなければならないことである．基底関数の数 M がデータ数 N より非常に小さい状況では，基底関数モデルを考える方が計算量的な観点からは都合が良い．

一方，ガウス過程で考えることは，無限個の基底関数でしか表せないような共分散関数を考えることができるという利点がある．

ガウス過程による回帰-最近 (?) の話題-

計算量の観点から，大きな訓練集合に対して，ガウス過程の直接的な適用は不可能であるため，さまざまな近似手法が提案されている．

また，複数の目標変数に対する拡張は容易であることが知られている．さらに，ガウス過程による回帰の拡張には，教師なし学習のための低次元の多様体上の分布や確率微分方程式の解のモデル化がある．

6.4.3 HP の学習

HP の学習-モチベーション-

ガウス過程による予測は，ある程度，共分散関数の選択に依存している．実際には，共分散関数をあらかじめ固定するよりも，パラメトリックな関数の族を考えて，そのパラメータをデータから推定する方が好まれる場合もある．これらのパラメータは通常のパラメトリックモデルにおける HP に対応しており，相関のスケールやノイズの精度などを調整する．

HP を学習する方法は，尤度関数 $p(\boldsymbol{t} \mid \boldsymbol{\theta})$ の評価に基づいている．ここで，$\boldsymbol{\theta}$ はガウス過程のモデルの HP とする．最も単純なアプローチは，対数尤度関数を最大化するような $\boldsymbol{\theta}$ の点推定を行うことである．対数尤度の最大化は，効率の良い，共役勾配法などの勾配を用いた最適化アルゴリズムが知られている．

HP の学習-対数尤度関数の最大化 (1/2)-

仮定：$\boldsymbol{C}_N$ の $\boldsymbol{\theta}$ に関する微分は簡単に求まる．

式 (6.61)[5] より，前節でみたガウス過程による回帰モデルにおける対数尤度関数は次式で与えられる．

$$\ln p(\boldsymbol{t} \mid \boldsymbol{\theta}) = -\frac{1}{2}\ln|\boldsymbol{C}_N| - \frac{1}{2}\boldsymbol{t}^\top \boldsymbol{C}_N^{-1}\boldsymbol{t} - \frac{N}{2}\ln(2\pi) \tag{6.69}$$

この関数を $\boldsymbol{\theta}$ で微分すると，次式が得られる（式 (C.21),(C.22)[6] を使う）．

$$\frac{\partial}{\partial \theta_i}\ln p(\boldsymbol{t} \mid \boldsymbol{\theta}) = -\frac{1}{2}\mathrm{Tr}\left(\boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_N}{\partial \theta_i}\right) + \frac{1}{2}\boldsymbol{t}^\top \boldsymbol{C}_N^{-1}\frac{\partial \boldsymbol{C}_N}{\partial \theta_i}\boldsymbol{C}_N^{-1}\boldsymbol{t} \tag{6.70}$$

一般的には，$\ln p(\boldsymbol{t} \mid \boldsymbol{\theta})$ は非凸であるため，複数の極大点を持ちうる．

---

[5] $p(\boldsymbol{t}) = \mathcal{N}(\boldsymbol{t} \mid \boldsymbol{0}, \boldsymbol{C}_N)$ (6.61)

[6] $\frac{\partial}{\partial x}(\boldsymbol{A}^{-1}) = -\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x}\boldsymbol{A}^{-1}$ (C.21), $\frac{\partial}{\partial x}\ln|\boldsymbol{A}| = \mathrm{Tr}(\boldsymbol{A}^{-1}\frac{\partial \boldsymbol{A}}{\partial x})$ (C.22)

HP の学習-対数尤度関数の最大化 (2/2)-

1 スライド前の手法とは別に，$\boldsymbol{\theta}$ の事前分布を定義して，対数事後分布を勾配法を用いて最大化することも考えられるが，完全にベイズ的な扱いをするには，事前分布 $p(\boldsymbol{\theta})$ と尤度関数 $p(\boldsymbol{t} \mid \boldsymbol{\theta})$ の積で重み付けされた $\boldsymbol{\theta}$ を周辺化したものを評価する必要がある．ところが，一般に，厳密な周辺化は不可能であるため，近似を用いる必要がある．

6.4.4 関連度自動決定

関連度自動決定-導入-

前節で，ガウス過程におけるパラメータの値を決定するために最尤推定を用いた．この方法は，各入力変数に対して別々のパラメータを与えるように拡張可能であり，入力間の相対的な重要度をデータから決定することに応用可能である．

本節では，ガウス過程での関連度自動決定 (ARD[7]) について 2 つの例を挙げる．（詳しい説明については，7.2.2 節）

（例１）2 次元の入力空間 $\boldsymbol{x} = (x_1, x_2)$ をもつガウス過程を考える．ここで，カーネル関数は次の形のものを用いる．

$$k(\boldsymbol{x}, \boldsymbol{x}') := \theta_0 \exp\left\{\frac{1}{2}\sum_{i=1}^{2}\eta_i(x_i - x_i')^2\right\} \tag{6.71}$$

[7] Automatic Relevance Determination

関連度自動決定-例 1-

図 6.9 にガウス仮定における ARD 事前分布からのサンプルを示す．

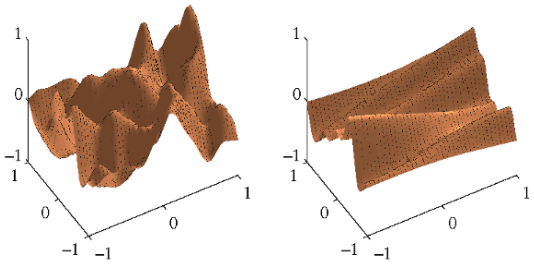


図 6.9 左：$\eta_1 = \eta_2 = 1$, 右：$\eta_1 = 1, \eta_2 = 0.01$

図 6.9 から，パラメータ η_i が小さくなると，関数の値が対応する入力変数 x_i の変化に対して，敏感でなくなることが分かる．最尤推定によって，これらのパラメータをデータに適応させると，対応する η_i の値が小さくなることから，予測分布にあまり寄与しない入力変数を検出することが可能になる．

関連度自動決定-例 2-

（例２）ARD を x_1, x_2, x_3 の 3 次元の入力変数を持つ単純な人工データに対して適用した結果について見る．（HP について最尤推定を行った結果は図 6.10）

- x_1: ガウス分布を用いて 100 個生成 $\longrightarrow$ 最重要
- x_2: 対応する x_1 の値にガウスノイズを付与 $\longrightarrow$ ワンチャン
- x_3: x_1, x_2 とは独立にガウス分布から生成 $\longrightarrow$ ゴミ
- 目標変数 t: x_1 に対して，$\sin(2\pi x_1)$ を適用して，さらにガウスノイズを付与

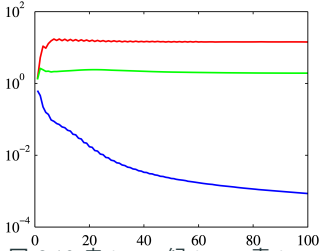


図 6.10 赤：η_1，緑：η_2，青：η_3

実際，η_1 は大きな値に，η_2 はまぁまぁの値に，η_3 は非常に小さい値に収束している．

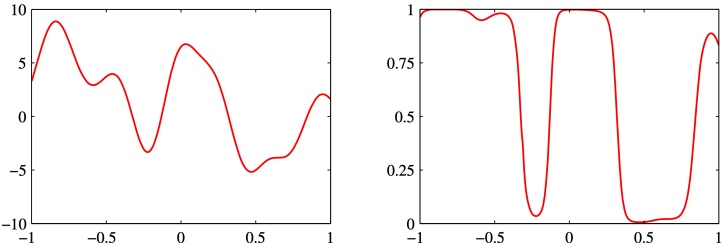
6.4.5 ガウス過程による分類

ガウス過程による分類-導入-

確率的なアプローチを用いた分類では，訓練データ集合が与えられたときに，新しい入力ベクトルに対する目標変数の事後確率をモデル化することが目的である．これらの確率は $[0, 1]$ に収まる必要があるが，ガウス過程のモデルの予測は実数値全体での値を取りうる．この問題を解決するために，ガウス過程の出力を適切な非線形関数の活性化関数を用いて変換することができる．

ガウス過程による分類-(0,1) への変換-

まず目標変数が $t \in \{0, 1\}$ であるような 2 クラス分類問題を考える．関数 $a(\boldsymbol{x})$ の上でのガウス過程を定義して，これをロジスティックシグモイド関数 $y = \sigma(a)$ で変換することで，$y \in (0, 1)$ であるような関数 $y(\boldsymbol{x})$ の上での非ガウス確率過程が得られる．図 6.11 は 1 次元の入力空間の場合の例である．



左：関数 $a(\boldsymbol{x})$ に対するガウス過程の事前分布からのサンプル，右：ロジスティックシグモイド関数での変換結果

ガウス過程による分類-事前分布-

目標変数 t の確率分布は次のベルヌーイ分布で与えられる．

$$p(t \mid a) = \sigma(a)^t (1 - \sigma(a))^{1-t} \tag{6.73}$$

要素 $a(\boldsymbol{x}_1), \ldots, a(\boldsymbol{x}_N), a(\boldsymbol{x}_{N+1})$ を持つベクトル $\boldsymbol{a}_{N+1}$ に対するガウス過程による事前分布は次式で与えられる．（式 (6.64) と同様）

$$p(\boldsymbol{a}_{N+1}) = \mathcal{N}(\boldsymbol{a}_{N+1} \mid \boldsymbol{0}, \boldsymbol{C}_{N+1}) \tag{6.74}$$

全ての訓練データ点は正しいクラスラベルが与えられていると仮定すると，回帰の場合と異なり，共分散行列はノイズ項を含まない．しかし，数値的な安定性から，共分散行列の正定値性を保証するために，パラメータ ν をもつノイズのような項を入れておくと便利である．このとき，共分散行列 $\boldsymbol{C}_{N+1}$ の各要素は次式で与えられる．

$$C_{ij} = k_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j) + \nu \delta_{ij} \tag{6.75}$$

k: $\boldsymbol{\theta}$ でパラメタライズされたカーネル関数，ν: 正数

ガウス過程による分類-予測分布の導出-

回帰タスクでは，予測分布 $p(t_{N+1} \mid \boldsymbol{t})$ を求めたが，2 クラス分類問題においては，$p(t_{N+1} = 0 \mid \boldsymbol{t}_N)$ は $1 - p(t_{N+1} = 1 \mid \boldsymbol{t}_N)$ によって与えられるため，$p(t_{N+1} = 1 \mid \boldsymbol{t}_N)$ を予測するだけで十分である．求めるべき予測分布は次式で与えられる．

$$p(t_{N+1} = 1 \mid \boldsymbol{t}_N) = \int p(t_{N+1} = 1 \mid a_{N+1}) p(a_{N+1} \mid \boldsymbol{t}_N) da_{N+1} \tag{6.76}$$

ここで，$p(t_{N+1} = 1 \mid a_{N+1}) = \sigma(a_{N+1})$ とする．

$p(a_{N+1} \mid \boldsymbol{t}_N)$ について，6.4.6 節で式変形を行う．

式 (6.76) の積分は解析的に求めることは不可能であるので近似を行う．（次ページ）

ガウス過程による分類-予測分布の積分の近似-

この積分は，解析的に求めることは不可能であるため，サンプリング (Neal, 1997) を用いて近似される．あるいは別の方法として，解析的な近似に基づくテクニックを用いることもできる．4.5.2 節において，ガウス分布によるロジスティックシグモイド関数の重畳積分の近似公式 (4.153) を示したが，この結果を用いて (6.76) の積分を評価し，事後分布 $p(a_{N+1}|\mathbf{t}_N)$ のガウス分布による近似を求めることができる．ガウス分布による近似は，通常，中心極限定理（☞2.3 節）によって，真の事後分布がデータ点の数の増加とともにガウス分布に近づくことから正当化される．ガウス過程の場合には，変数の数はデータ点の数の増加とともに大きくなるため，この議論は直接には適用されない．しかしながら，入力 $\mathbf{x}$ の空間のある決まった領域に含まれるデータ点の数が増加すると考えると，対応する関数 $a(\mathbf{x})$ の不確定性は減少し，やはり結果として，漸近的にガウス分布へと近づくことになる (Williams and Barber, 1998).

ガウス分布による近似の方法としては，3 つの異なるアプローチが提案されている．1 つ目は，**変分推論法** (variational inference) （☞10.1 節）に基づく方法 (Gibbs and MacKay, 2000) で，ロジスティックシグモイド関数の局所的な変分近似 (10.144) を用いる．この方法では，シグモイド関数の積を，ガウス分布の積によって近似し，それによって，$\mathbf{a}_N$ の周辺化を解析的に行うことができる．この方法では，尤度関数 $p(\mathbf{t}_N|\boldsymbol{\theta})$ の下界を得ることもできる．ガウス過程による分類への変分アプローチの枠組みは多クラス ($K > 2$) の分類問題の場合にも拡張可能であり，ソフトマックス関数をガウス分布によって近似することによって達成される (Gibbs, 1997).

2 つ目のアプローチは，**EP 法** (expectation propagation method) （☞10.7 節）を用いたものである (Opper and Winther, 2000b; Minka, 2001b; Seeger, 2003). 真の事後分布は単峰性を持つため，後で見るように，EP 法は良い結果をもたらす．

6.4.6 ラプラス近似

ラプラス近似-ベイズの定理の適用-

ガウス仮定による分類に対する 3 つ目のアプローチは，ラプラス近似（4.4 節）を用いた方法である．

ベイズの定理と $p(\boldsymbol{t}_N \mid a_{N+1}, \boldsymbol{a}_N) = p(\boldsymbol{t}_N \mid \boldsymbol{a}_N)$ を用いることで，次式が得られる．

$$\begin{aligned} p(a_{N+1}, \boldsymbol{a}_N \mid \boldsymbol{t}_N) &= \frac{p(a_{N+1}, \boldsymbol{a}_N) p(\boldsymbol{t}_N \mid a_{N+1}, \boldsymbol{a}_N)}{p(\boldsymbol{t}_N)} \\ &= \frac{p(a_{N+1} \mid \boldsymbol{a}_N) p(\boldsymbol{a}_N) p(\boldsymbol{t}_N \mid \boldsymbol{a}_N)}{p(\boldsymbol{t}_N)} \\ &= p(a_{N+1} \mid \boldsymbol{a}_N) p(\boldsymbol{a}_N \mid \boldsymbol{t}_N) \end{aligned}$$

よって，次式が成り立つ．

$$p(a_{N+1} \mid \boldsymbol{t}_N) = \int p(a_{N+1}, \boldsymbol{a}_N \mid \boldsymbol{t}_N) d\boldsymbol{a}_N = \int p(a_{N+1} \mid \boldsymbol{a}_N) p(\boldsymbol{a}_N \mid \boldsymbol{t}_N) d\boldsymbol{a}_N \quad (6.77)$$

ラプラス近似-データについての項の計算-

条件付き分布 $p(a_{N+1} \mid \boldsymbol{a}_N)$ は式 (6.66),(6.67) を用いて（$\boldsymbol{t}$ と $\boldsymbol{a}_N$ を入れ替えて），次式で与えられる．

$$p(a_{N+1} \mid \boldsymbol{a}_N) = \mathcal{N}(a_{N+1} \mid \boldsymbol{k}^\top \boldsymbol{C}_N^{-1} \boldsymbol{a}_N, c - \boldsymbol{k}^\top \boldsymbol{C}_N^{-1} \boldsymbol{k}) \tag{6.78}$$

また，式 (6.73)[8] より，データ点が互いに独立であると仮定すると，次式が成り立つ．

$$\begin{aligned} p(\boldsymbol{t}_N \mid \boldsymbol{a}_N) &= \prod_{n=1}^{N} \sigma(a_n)^{t_n} (1 - \sigma(a_n))^{1-t_n} = \prod_{n=1}^{N} \sigma(a_n)^{t_n} \left(\sigma(-a_n)\right)^{1-t_n} \\ &= \prod_{n=1}^{N} \left(\frac{\sigma(a_n)}{\sigma(-a_n)}\right)^{t_n} \sigma(-a_n) = \prod_{n=1}^{N} e^{a_n t_n} \sigma(-a_n) \end{aligned} \tag{6.79}$$

[8] $p(t \mid a) = \sigma(a)^t (1 - \sigma(a))^{1-t}$

ラプラス近似-事後分布の導出-

事後分布 $p(\boldsymbol{a}_N \mid \boldsymbol{t}_N)$ の対数は次式で与えられる．

$$\begin{aligned}\ln p(\boldsymbol{a}_N \mid \boldsymbol{t}_N) &= \ln \frac{p(\boldsymbol{a}_N)p(\boldsymbol{t}_N \mid \boldsymbol{a}_N)}{p(\boldsymbol{t}_N)} \\ &= \ln p(\boldsymbol{a}_N) + \ln p(\boldsymbol{t}_N \mid \boldsymbol{a}_N) - \ln p(\boldsymbol{t}_N)\end{aligned}$$

等式の右辺について，$\boldsymbol{a}_N$ に依存しない項を除いたものを，$\Psi(\boldsymbol{a}_N)$ とする．このとき，式 (6.79)[9] と $p(\boldsymbol{a}_N) = \mathcal{N}(\boldsymbol{a}_N \mid \boldsymbol{0}, \boldsymbol{C}_N)$ から，次式が成立する．

$$\begin{aligned}\Psi(\boldsymbol{a}_N) &= \ln p(\boldsymbol{a}_N) + \ln p(\boldsymbol{t}_N \mid \boldsymbol{a}_N) \\ &= -\frac{1}{2}\boldsymbol{a}_N^\top \boldsymbol{C}_N^{-1}\boldsymbol{a}_N - \frac{N}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{C}_N| + \boldsymbol{t}_N^\top \boldsymbol{a}_N - \sum_{n=1}^{N}\ln(1+e^{a_n}) \quad (6.80)\end{aligned}$$

[9] $p(\boldsymbol{t}_N \mid \boldsymbol{a}_N) = \prod_{n=1}^{N} e^{a_n t_n}\sigma(-a_n)$ (6.79)

ラプラス近似-事後分布のモードの導出 (1/2)-

次に，$\Psi(\boldsymbol{a}_N)$ についてラプラス近似を行う．まず，事後分布のモードを求めたい．それには，$\Psi(\boldsymbol{a}_N)$ の勾配が必要であるが，これは次式で与えられる．

$$\nabla\Psi(\boldsymbol{a}_N) = \boldsymbol{t}_N - \boldsymbol{\sigma}_N - \boldsymbol{C}_N^{-1}\boldsymbol{a}_N \tag{6.81}$$

ここで，$\boldsymbol{\sigma}_N \in \mathbb{R}^N$ は要素 $\sigma(a_n)$ を持つベクトルである．勾配を $\mathbf{0}$ として，モードを求めることは不可能であるので，ニュートンーラフソン法に基づく繰り返し法で求める．これは結果的に反復再重み付け最小二乗法（4.3.3 節）となる．その計算には，Ψ の二階微分が必要となるが，これは次式で与えられる．

$$\nabla\nabla\Psi(\boldsymbol{a}_N) = -\boldsymbol{W}_N - \boldsymbol{C}_N^{-1} \tag{6.82}$$

ここで，$\boldsymbol{W}_N \coloneqq \mathrm{diag}(\sigma(a_1)(1-\sigma(a_1)), \ldots, \sigma(a_N)(1-\sigma(a_N)))$ である．（式 (4.88)[10] を用いた．）

[10] $\frac{d\sigma}{da} = \sigma(1-\sigma)$

ラプラス近似-事後分布のモードの導出 (2/2)-

対角行列 $\boldsymbol{W}_N$ の対角要素 $\sigma(a_n)(1-\sigma(a_n)) \in (0, 1/4)$ であるので，$\boldsymbol{W}_N$ は正定値行列である．また，$\boldsymbol{C}_N$ は式 (6.75)[11] より，正定値行列であり，その逆行列も正定値行列である．正定値行列同士の和は正定値行列であるので，$\nabla\nabla\Psi(\boldsymbol{a}_N)$ は負定値行列であるため，事後分布の対数 $\ln p(\boldsymbol{a}_N \mid \boldsymbol{t}_N)$ は凹関数である．したがって，事後分布 $p(\boldsymbol{a}_N \mid \boldsymbol{t}_N)$ は凹関数となり，大域的な最大解を持つ．しかし，これは正規分布ではない．

[11] $C_{ij} = k_{\boldsymbol{\theta}}(\boldsymbol{x}_i, \boldsymbol{x}_j) + \nu\delta_{ij} \quad (6.75)$

ラプラス近似-ニュートンーラフソン法によるモードとヘッセ行列の計算-

ニュートンーラフソンの公式 (4.92)[12] から，$\boldsymbol{a}_N$ の逐次更新式は次式で与えられる．((6.81), (6.82)[13] を代入するだけ)

$$\boldsymbol{a}_N^{\text{new}} = \boldsymbol{C}_N(\boldsymbol{I} + \boldsymbol{W}_N\boldsymbol{C}_N)^{-1}\{\boldsymbol{t}_N - \boldsymbol{\sigma}_N + \boldsymbol{W}_N\boldsymbol{a}_N\} \tag{6.83}$$

反復はモード $\boldsymbol{a}^\star$ に収束するまで続く．モードにおいては勾配 $\nabla\Psi(\boldsymbol{a}_N)$ は $\boldsymbol{0}$ となるため，$\boldsymbol{a}^\star$ は次式を満たす．

$$\boldsymbol{a}_N^\star = \boldsymbol{C}_N(\boldsymbol{t}_N - \boldsymbol{\sigma}_N) \tag{6.84}$$

事後分布のモード $\boldsymbol{a}^\star$ に到達したら，ヘッセ行列を次のように求める．

$$\boldsymbol{H} \coloneqq -\nabla\nabla\Psi(\boldsymbol{a}_N^\star) = \boldsymbol{W}_N(\boldsymbol{a}_N^\star) + \boldsymbol{C}_N^{-1} \tag{6.85}$$

---

[12] $\boldsymbol{w}^{\text{new}} = \boldsymbol{w}^{\text{old}} - \boldsymbol{H}^{-1}\nabla\boldsymbol{E}(\boldsymbol{w})$ $(\boldsymbol{H}\colon \nabla\nabla E(\boldsymbol{w}))$ (4.92)

[13] $\nabla\Psi(\boldsymbol{a}_N) = \boldsymbol{t}_N - \boldsymbol{\sigma}_N - \boldsymbol{C}_N^{-1}\boldsymbol{a}_N$ (6.81), $\nabla\nabla\Psi(\boldsymbol{a}_N) = -\boldsymbol{W}_N - \boldsymbol{C}_N^{-1}$ (6.82)

ラプラス近似-ニュートンーラフソン法によるガウス分布の近似-

前スライドの計算により，事後分布 $p(\boldsymbol{a}_N \mid \boldsymbol{t}_N)$ のガウス分布による近似は次のように求まる．

$$p(\boldsymbol{a}_N \mid \boldsymbol{t}_N) \approx q(\boldsymbol{a}_N) = \mathcal{N}(\boldsymbol{a}_N \mid \boldsymbol{a}_N^\star, \boldsymbol{H}^{-1}) \tag{6.86}$$

これと式 (6.78)[14] を組み合わせることで，式 (6.77)[15] の積分を評価できる．なお，式 (2.115)[16] の結果を用いると，次式が得られる．（ゴリゴリ計算）

$$\mathbb{E}[a_{N+1} \mid \boldsymbol{t}_N] = \boldsymbol{k}^\top (\boldsymbol{t}_N - \boldsymbol{\sigma}_N) \tag{6.87}$$

$$\mathrm{var}[a_{N+1} \mid \boldsymbol{t}_N] = c - \boldsymbol{k}^\top (\boldsymbol{W}_N^{-1} + \boldsymbol{C}_N)^{-1} \boldsymbol{k} \tag{6.88}$$

すなわち，

$$p(a_{N+1} \mid \boldsymbol{t}_N) = \mathcal{N}(a_{N+1} \mid \boldsymbol{k}^\top (\boldsymbol{t}_N - \boldsymbol{\sigma}_N), c - \boldsymbol{k}^\top (\boldsymbol{W}_N^{-1} + \boldsymbol{C}_N)^{-1} \boldsymbol{k})$$

[14] $p(a_{N+1} \mid \boldsymbol{a}_N) = \mathcal{N}(a_{N+1} \mid \boldsymbol{k}^\top \boldsymbol{C}_N^{-1} \boldsymbol{a}_N, c - \boldsymbol{k}^\top \boldsymbol{C}_N^{-1} \boldsymbol{k})$ (6.78)

[15] $p(a_{N+1} \mid \boldsymbol{t}_N) = \int p(a_{N+1} \mid \boldsymbol{a}_N) p(\boldsymbol{a}_N \mid \boldsymbol{t}_N) d\boldsymbol{a}_N$ (6.77)

[16] $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}), p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{L}^{-1}) \implies p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y} \mid \boldsymbol{A}\boldsymbol{\mu} + \boldsymbol{b}, \boldsymbol{L}^{-1} + \boldsymbol{A}\boldsymbol{\Lambda}^{-1}\boldsymbol{A}^\top)$ (2.115)

ラプラス近似-予測分布の積分の近似-

以上の議論から，$p(a_{N+1} \mid \boldsymbol{t}_N)$ のガウス分布による近似が求まったため，式 (4.153)[17] を用いて，予測分布の式 (6.76)[18] の積分を評価できる．

さらに，共分散関数のパラメータ $\boldsymbol{\theta}$ を決定する必要がある．ここでは，尤度関数 $p(\boldsymbol{t}_N \mid \boldsymbol{\theta})$ を最大化することで $\boldsymbol{\theta}$ を決定する．

尤度関数は次式で与えられる．

$$p(\boldsymbol{t}_N \mid \boldsymbol{\theta}) = \int p(\boldsymbol{t}_N \mid \boldsymbol{a}_N) p(\boldsymbol{a}_N \mid \boldsymbol{\theta}) d\boldsymbol{a}_N \tag{6.89}$$

この積分は解析的には求められないため，再びラプラス近似を行う．

[17] $\int \sigma(a)\mathcal{N}(a \mid \mu, \sigma^2)da \approx \sigma(\kappa(\sigma^2)\mu) \quad (4.153), \qquad \kappa(\sigma^2) = (1 + \pi\sigma^2/8)^{-1/2} \quad (4.154)$

[18] $p(t_{N+1} = 1 \mid \boldsymbol{t}_N) = \int p(t_{N+1} = 1 \mid a_{N+1}) p(a_{N+1} \mid \boldsymbol{t}_N) da_{N+1} \quad (6.76)$

ラプラス近似-カーネル関数の HP の最適化-

$$p(\boldsymbol{t}_N \mid \boldsymbol{\theta}) = \int p(\boldsymbol{t}_N \mid \boldsymbol{a}_N) p(\boldsymbol{a}_N \mid \boldsymbol{\theta}) d\boldsymbol{a}_N \tag{6.89}$$

について，式 (4.135)[19] を利用すると，次のように対数尤度関数の近似を求めることができる．$(\boldsymbol{Z} \coloneqq p(\boldsymbol{t}_N \mid \boldsymbol{\theta}), f(\boldsymbol{z}) \coloneqq p(\boldsymbol{t}_N \mid \boldsymbol{a}_N) p(\boldsymbol{a}_N \mid \boldsymbol{\theta}), \boldsymbol{z} \coloneqq \boldsymbol{a}_N)$

$$\begin{aligned} p(\boldsymbol{t}_N \mid \boldsymbol{\theta}) &\approx p(\boldsymbol{t}_N \mid \boldsymbol{a}_N^\star) p(\boldsymbol{a}_N^\star \mid \boldsymbol{\theta}) \frac{(2\pi)^{N/2}}{|\boldsymbol{H}|^{1/2}} \\ \ln p(\boldsymbol{t}_N \mid \boldsymbol{\theta}) &\approx \ln p(\boldsymbol{t}_N \mid \boldsymbol{a}_N^\star) + \ln p(\boldsymbol{a}_N^\star \mid \boldsymbol{\theta}) - \frac{1}{2} \ln |\boldsymbol{H}| + \frac{N}{2} \ln(2\pi) \\ &= \ln p(\boldsymbol{t}_N \mid \boldsymbol{a}_N^\star) + \ln p(\boldsymbol{a}_N^\star \mid \boldsymbol{\theta}) - \frac{1}{2} \ln |\boldsymbol{W}_N(\boldsymbol{a}_N^\star) + \boldsymbol{C}_N^{-1}| + \frac{N}{2} \ln(2\pi) \end{aligned} \tag{6.90}$$

[19] $\boldsymbol{Z} = \int f(\boldsymbol{z}) d\boldsymbol{z} \approx f(\boldsymbol{z_0}) \frac{(2\pi)^{M/2}}{|\boldsymbol{A}|^{1/2}}$ (f: ガウス関数, $\boldsymbol{z}_0$: モード, $\boldsymbol{A} \coloneqq -\nabla\nabla \ln f(\boldsymbol{z})|_{\boldsymbol{z}=\boldsymbol{z}_0}$: ヘッセ行列)(4.135)

ラプラス近似-カーネル関数の HP の最適化-

式 (6.90) の θ に関する勾配は添付資料のようにして求めることができる．（とても大変）この勾配を基に，対数尤度関数の最大化を行えばよい．

6.4.7 NN との関係

すでに，ニューラルネットワークによって表現できる関数の種類は，隠れユニットの数 M に依存し，十分に大きい M を取ることによって，2 層のニューラルネットワークは任意の関数を任意の精度で近似できることを見てきたが，最尤推定の枠組みでは，過学習を避けるために，隠れユニットの数は（訓練集合のサイズに合う程度まで）制限する必要がある．しかしながら，ベイズの観点からは，訓練集合のサイズに依存して，ネットワークのパラメータの数を制限することは，ほとんど意味を持たない．

ベイズニューラルネットワークでは，パラメータベクトル $\mathbf{w}$ の事前分布とネットワーク関数 $f(\mathbf{x}, \mathbf{w})$ を組み合わせることによって，$y(\mathbf{x})$ の上の関数についての事前分布が得られる．ここで，$\mathbf{y}$ はネットワークの出力ベクトルである．Neal (1996) では，$\mathbf{w}$ の事前分布として広いクラスの分布に対して，ニューラルネットワークによって生成される関数の分布が，$M \to \infty$ の極限においてガウス過程に近づくことが示されている．しかしながら，この極限においては，ニューラルネットワークの出力変数は独立になることに注意せねばならない．ニューラルネットワークで最も有用な点の 1 つは，出力が隠れユニットを共有することであり，これによって，お互いに「統計的な強度を借りる」ことが可能になる．つまり，それぞれの隠れユニットに関連付けられた重みは，(1 つではなく）すべての出力変数から影響を受けることになる．この性質は，極限でのガウス過程においては失われてしまう．

すでに，ガウス過程は，その共分散（カーネル）関数によって決定されることを見てきたが，Williams (1998) では，プロビット関数とガウス関数の 2 つの活性化関数を隠れユニットに用いた場合の共分散を明示的に導いている．零を中心としたガウス関数による重みの事前分布は，重み空間における平行移動不変性が成り立たないため，結果として，これらのカーネル関数 $k(\mathbf{x}, \mathbf{x}')$ は不変にはならない，つまり，差 $\mathbf{x} - \mathbf{x}'$ の関数として表現されない．

共分散関数を直接的に扱うことによって，事前分布の重みの分布を暗黙的に周辺化していることになる．事前分布の重みパラメータが，超パラメータによって決定されるならば，図 5.11 の無限個の隠れユニットの例からもわかるように，それらの値は，関数の分布の長さスケールを決定する．なお，超パラメータについては解析的に周辺化することはできないため，6.4 節で見たようなテクニックを用いる必要がある．