PRML の 3 章のための数学

サイボウズ・ラボ 光成滋生

2011 年 4 月 27 日

1　概要

この文章は『パターン認識と機械学習』（以下 PRML）の 3 章を理解するために必要な数学をまとめてみたものです．間違い, 質問などございましたら `herumi@nifty.com` または `twitterID:herumi` までご連絡ください.

2　最小二乗法

2.1　微分の復習

$\boldsymbol{x}$, $\boldsymbol{y}$ を縦ベクトルとして

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{y}) = \boldsymbol{y}.$$

$$\frac{\partial}{\partial \boldsymbol{y}}(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{y}) = \boldsymbol{x}.$$

ここで $\frac{\partial}{\partial \boldsymbol{x}}$ は $\frac{\partial}{\partial x_i}$ を縦に並べた縦ベクトルとする．2 章でも述べたが $\frac{\partial}{\partial \boldsymbol{x}}$ を ∇ と書くこともあるが PRML では場所によって縦ベクトル（3.22）だったり, 横ベクトル（3.13）だったりする．常に縦ベクトルとしたほうが混乱は少ない.

2.2　誤差関数の最小化

$$f(\boldsymbol{w}) = \sum_{n=1}^{N}\{t_n - \boldsymbol{w}^{\boldsymbol{T}}\phi(\boldsymbol{x}_n)\}^2 + \lambda \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{w}$$

とする．ここで $\boldsymbol{w}$ と $\phi(\boldsymbol{x}_n)$ は M 次元縦ベクトルである.

$$\Phi^{\boldsymbol{T}} = (\phi(\boldsymbol{x}_1)\cdots\phi(\boldsymbol{x}_N))$$

とおく．Φ は N 行 M 列の行列である．$f(\boldsymbol{w})$ を w で微分しよう.

$$\frac{\partial}{\partial \boldsymbol{w}}f(\boldsymbol{w}) = 2\sum_{n=1}^{N}\{t_n - \boldsymbol{w}^{\boldsymbol{T}}\phi(\boldsymbol{x}_n)\}(-\phi(\boldsymbol{x}_n)) + 2\lambda\boldsymbol{w}.$$

一般に縦ベクトル $\boldsymbol{x}$, $\boldsymbol{y}$ に対して

$$(\boldsymbol{x}^{\boldsymbol{T}}\boldsymbol{y})\boldsymbol{y} = (\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{x})\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{x}) = (\boldsymbol{y}\boldsymbol{y}^{\boldsymbol{T}})\boldsymbol{x}$$

だから $\boldsymbol{t} = (t_1, \ldots, t_N)^{\boldsymbol{T}}$ とおくと

$$\begin{aligned}\frac{1}{2}\frac{\partial}{\partial \boldsymbol{w}} f(\boldsymbol{w}) &= -\sum_n t_n \phi(\boldsymbol{x}_n) + \sum_n (\phi(\boldsymbol{x}_n)\phi(\boldsymbol{x}_n)^{\boldsymbol{T}})\boldsymbol{w} + \lambda \boldsymbol{w} \\ &= -\Phi^{\boldsymbol{T}}\boldsymbol{t} + \Phi^{\boldsymbol{T}}\Phi\boldsymbol{w} + \lambda\boldsymbol{w} \\ &= -\Phi^{\boldsymbol{T}}\boldsymbol{t} + (\Phi^{\boldsymbol{T}}\Phi + \lambda I)\boldsymbol{w} = 0.\end{aligned}$$

よって $\det(\lambda I + \Phi^{\boldsymbol{T}}\Phi) \neq 0$ のとき

$$\boldsymbol{w}_{\mathrm{ML}} = (\lambda I + \Phi^{\boldsymbol{T}}\Phi)^{-1}\Phi^{\boldsymbol{T}}\boldsymbol{t}$$

が最尤解. $\boldsymbol{y} = \Phi\boldsymbol{w}$ が予測値である.

2.3 正射影

前節で $\lambda = 0$ のときを考える.

$$\boldsymbol{y} = \Phi(\Phi^{\boldsymbol{T}}\Phi)^{-1}\Phi^{\boldsymbol{T}}\boldsymbol{t}$$

となる. ここでこの式の幾何学的な解釈を考えてみよう.

$\Phi = (a_1 \cdots a_M)$ と縦ベクトルの集まりで表す. $N - M$ 個のベクトル $b_1, \ldots, b_{N-M}$ を追加して, $\{a_1, \ldots, a_N, b_1, \ldots, b_{N-M}\}$ 全体で N 次元ベクトル空間の基底であるようにとる. その際 b_i を a_j と直交するようにとれる.

$$a_i^{\boldsymbol{T}} b_j = 0.$$

さて $X = \Phi(\Phi^{\boldsymbol{T}}\Phi)^{-1}\Phi^{\boldsymbol{T}}$ とおくと, $X\Phi = \Phi$. これは $Xa_i = a_i$ を意味する. つまり X は $a_1, \ldots, a_M$ で生成される部分空間 V=$\langle a_1, \ldots, a_M\rangle$ の点を動かさない. また b_j のとりかたから $Xb_j = 0$ も成り立つ. つまり X は部分空間 $\langle b_1, \ldots, b_{N-M}\rangle$ の点を 0 につぶす.

二つ合わせると, X は任意の点を部分空間 V 方向につぶす写像, つまり V への正射影写像と解釈できる. 式で書くと任意の点 $\boldsymbol{t}$ を $\boldsymbol{t} = \sum_i s_i a_i + \sum_i t_i b_j$ と表したとすると,

$$\boldsymbol{y} = X\boldsymbol{t} = \sum_i s_i a_i$$

となる. $\boldsymbol{t}$ から $\boldsymbol{y}$ への変換を係数だけを使って書いてみると

$$X : (s_1, \ldots, s_N, t_1, \ldots, t_{N-M}) \to (s_1, \ldots, s_N, 0, \ldots, 0).$$

これを見ると正射影のニュアンスがより明確になる.

2.4 行列での微分

x を n 次元ベクトル, A を m 行 n 列として $y = Ax$ とおく.

$$f(A) = ||y||^2 = (Ax)^{\boldsymbol{T}} Ax$$

を A で微分してみよう.

$$(Ax)^{\boldsymbol{T}}Ax = \sum_s (Ax)_s(Ax)_s = \sum_s(\sum_t a_{st}x_t)(\sum_u a_{su}x_u) = \sum_{s,t,u} x_t x_u a_{st} a_{su}.$$

よって

$$\begin{aligned}
\frac{\partial}{\partial a_{ij}} f(A) &= \sum_{s,t,u} x_t x_u ((\frac{\partial}{\partial a_{ij}} a_{st})a_{su} + a_{st}\frac{\partial}{\partial a_{ij}} a_{su}) \\
&= \sum_{s,t,u} x_t x_u (\delta_{is}\delta_{jt} a_{su} + a_{st}\delta_{is}\delta_{ju}) \\
&= (\sum_u x_j x_u a_{iu}) + (\sum_t x_t x_j a_{it}) \\
&= 2\sum_u x_j x_u a_{iu} \\
&= 2x_j(Ax)_i \\
&= 2(Axx^{\boldsymbol{T}})_{ij}.
\end{aligned}$$

よって

$$\frac{\partial}{\partial A}||Ax||^2 = 2Axx^{\boldsymbol{T}}.$$

2.5 Woodbury の逆行列の公式

行列 A, B, C, D について

$$(A+BD^{-1}C)^{-1} = A^{-1} - A^{-1}B(D+CA^{-1}B)^{-1}CA^{-1}$$

が成り立つ.

（証明）

$$\begin{aligned}
A^{-1}B(D+CA^{-1}B)^{-1}CA^{-1} &= A^{-1}B((DB^{-1}A+C)A^{-1}B)^{-1}CA^{-1} \\
&= (DB^{-1}A+C)^{-1}CA^{-1} \\
&= (DB^{-1}(A+BD^{-1}C))^{-1}CA^{-1} \\
&= (A+BD^{-1}C)^{-1}BD^{-1}CA^{-1}
\end{aligned}$$

よって

$$\begin{aligned}
\text{左辺} &= (I-(A+BD^{-1}C)^{-1}BD^{-1}C)A^{-1} \\
&= (A+BD^{-1}C)^{-1}((A+BD^{-1}C)-BD^{-1}C)A^{-1} \\
&= \text{右辺}.
\end{aligned}$$

特に, A が n 次正方行列で B を n 次縦ベクトル $\boldsymbol{x}$, $C=\boldsymbol{x}^{\boldsymbol{T}}$, D を n 次単位行列とすると

$$(A+\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}})^{-1} = A^{-1} - \frac{(A^{-1}\boldsymbol{x})(\boldsymbol{x}^{\boldsymbol{T}}A^{-1})}{1+\boldsymbol{x}^{\boldsymbol{T}}A^{-1}\boldsymbol{x}} \tag{1}$$

が成り立つ.

2.6　正定値対称行列

n 次元実対称行列 A はある直行行列 P を用いて常に対角化可能であった.

$$P^{-1}AP = \mathrm{diag}(\lambda_1, \cdots, \lambda_n).$$

全ての固有値が正であるとき A を正定値といい, $A > 0$ とかく. 全ての固有値が正または 0 であるとき, 半正定値といい, $A \geq 0$ とかく.

任意の実ベクトル $\boldsymbol{x}$ について $\boldsymbol{y} = P\boldsymbol{x}$ とおくと $\boldsymbol{x}$ が $\mathbb{R}^n$ の全ての点をとるとき $\boldsymbol{y}$ も全ての点を渡る.

$$\boldsymbol{x}^{\boldsymbol{T}} A\boldsymbol{x} = \sum_i \lambda_i y_i^2$$

なので $A \geq 0$ ならば $\boldsymbol{x}^{\boldsymbol{T}} A\boldsymbol{x} \geq 0$. $A > 0$ のときは等号が成り立つのは $\boldsymbol{x} = 0$ のときのみである.

逆に任意の $\boldsymbol{x}$ について $\boldsymbol{x}^{\boldsymbol{T}} A\boldsymbol{x} \geq 0$ とすると, $\boldsymbol{y}$ として単位ベクトル $\boldsymbol{e_i}$ を考えれば $\lambda_i \geq 0$. つまり $A \geq 0$. 更に等号は $\boldsymbol{x} = 0$ のときに限るためには $\lambda_i > 0$. つまり $A > 0$ であることが分かる. まとめると

$$A \geq 0 \iff \lambda_i \geq 0 \text{ for } \forall i.$$

$$A > 0 \iff \lambda_i > 0 \text{ for } \forall i.$$

この同値性から $A > 0$ のとき $A^{-1} > 0$ も分かる. 定義から $A > 0$, $B > 0$ なら $A + B > 0$ も成り立つ.

また実ベクトル $\boldsymbol{v}$ に対して $A = \boldsymbol{v}\boldsymbol{v}^{\boldsymbol{T}}$ とおくと, A は実対称であり, 任意の $\boldsymbol{x}$ に対して

$$\boldsymbol{x}^{\boldsymbol{T}} A\boldsymbol{x} = (\boldsymbol{v}^{\boldsymbol{T}}\boldsymbol{x})^2 \geq 0$$

なので $A \geq 0$.

2.7　予測分布の分散

$S_N^{-1} = S_0^{-1} + \beta\Phi_N^{\boldsymbol{T}}\Phi_N$ としたときの予測分布の分散

$$\sigma_N^2 = \frac{1}{\beta} + \phi^{\boldsymbol{T}} S_N \phi$$

を考える. $\beta > 0$ であり, S_0 は共分散行列なので実正定値であることに注意する. まず計画行列 Φ_N は N が一つ増える毎に 1 行増える. $\boldsymbol{v}_N$（煩雑なので v と略記する）を M 次元縦ベクトルとして

$$\Phi_{N+1}^{\boldsymbol{T}} = \begin{pmatrix}\Phi_N^{\boldsymbol{T}} & v\end{pmatrix}$$

としよう. すると

$$S_{N+1}^{-1} = S_0^{-1} + \beta(\Phi_N^{\boldsymbol{T}}\Phi_N + vv^{\boldsymbol{T}}) = S_N^{-1} + \beta vv^{\boldsymbol{T}}.$$

行列 $\beta vv^{\boldsymbol{T}}$ は正定値であり, S_N に関して帰納法を使うと全ての S_N は正定値であることが分かる.

公式 1 を使って

$$\begin{aligned}
\sigma_{N+1}^2 &= \frac{1}{\beta} + \phi^{\boldsymbol{T}}(S_N^{-1} + \beta vv^{\boldsymbol{T}})^{-1}\phi \\
&= \frac{1}{\beta} + \phi^{\boldsymbol{T}}(S_N - \frac{(S_N v)(v^{\boldsymbol{T}} S_N)}{1 + v^{\boldsymbol{T}} S_N v})\phi \\
&= \sigma_N^2 - z
\end{aligned}$$

ここで S_N は対称なので

$$
\begin{aligned}
z &= \phi^{\boldsymbol{T}} \frac{(S_N v)(v^{\boldsymbol{T}} S_N)}{1 + v^{\boldsymbol{T}} S_N v} \phi \\
&= \frac{1}{1 + v^{\boldsymbol{T}} S_N v} (v^{\boldsymbol{T}} S_N \phi)^2.
\end{aligned}
$$

S_N は正定値なので任意の v に対して $v^{\boldsymbol{T}} S_N v \geq 0$. よって $z \geq 0$ となり

$$\sigma_{N+1}^2 \leq \sigma_N^2.$$

帰納法の流れを見ると,

$$\Phi_N^{\boldsymbol{T}} = \left(\boldsymbol{v}_1 \cdots \boldsymbol{v}_N\right)$$

とおくと

$$S_N^{-1} = S_0^{-1} + \beta \sum_{i=1}^{N} \boldsymbol{v}_i \boldsymbol{v}_i^{\boldsymbol{T}}$$

となることがわかる. $\boldsymbol{v}_i$ が基底関数のベクトルに訓練データの値を代入したものであることを考えると, 0 ベクトルになることは殆ど無い. また $N \to \infty$ で 0 になるわけでもない. つまりそれらの和はどんどん大きくなる. そういう状況の元では $\phi^{\boldsymbol{T}} S_N \phi$ は 0 に近づき,

$$\sigma_N^2 \to \frac{1}{\beta}$$

となる.

2.8 カルバック距離

$p(x)$, $q(x)$ を恒等的に 0 ではない確率密度関数とする. つまり $p(x), q(x) \geq 0$.

$$\mathrm{KL}(p||q) = \int p(x) \log \frac{p(x)}{q(x)} \, dx$$

をカルバック距離 (Kullback-Leibler 距離, 相対エントロピー) という.

距離といいつつ, $\mathrm{KL}(p||q) = \mathrm{KL}(q||p)$ とは限らないので距離の公理は満たさない. しかし, $\mathrm{KL}(p||q) \geq 0$ であり, $\mathrm{KL}(p||q) = 0 \iff p = q$ はいえる. これを示そう.

まず $S(x) = e^{-x} + x - 1$ について $S(x) \geq 0$ であり, $S(x) = 0 \iff x = 0$ である.

なぜなら $S'(x) = -e^{-x} + 1$. $S''(x) = e^{-x} \geq 0$ なので $S'(x)$ は単調増加. $S'(0) = 0$ より $x > 0$ なら $S'(x) > 0$, $x < 0$ なら $S'(x) < 0$. つまり $S(x)$ は 0 で最小値 0 をとる.

$$\int p(x) S(\log \frac{p(x)}{q(x)}) \, dx = \int p(x) (\frac{q(x)}{p(x)} + \log \frac{p(x)}{q(x)} - 1) \, dx = \mathrm{KL}(p||q) + \int (q(x) - p(x)) \, dx = \mathrm{KL}(p||q).$$

ここで p, q が確率密度関数なので $\int p(x) \, dx = 1$, $\int q(x) \, dx = 1$ であることを使った.

この式の左辺の被積分関数は常に 0 以上. よって $\mathrm{KL}(p||q) \geq 0$.

$\mathrm{KL}(p||q) = 0$ ならば殆ど全ての x について

$$p(x) S(\log \frac{p(x)}{q(x)}) = 0.$$

$p=0$ ではないので殆ど全ての x について

$$S(\log \frac{p(x)}{q(x)}) = 0.$$

$S(x)=0$ となる x は 0 のときだけだから, 殆ど全ての x について $p(x)=q(x)$.

真のモデル $p(D|M)$ があったときに, モデルエビデンス $p(D|M')$ とのカルバック距離 $\mathrm{KL}(p(D|M)||p(D|M'))$ は, 0 に近いほど真のモデルに近そうだということにする.

2.9 エビデンス関数の評価の式変形

$A=\alpha I+\beta\Phi^{\boldsymbol{T}}\Phi$ とおくと

$$\begin{aligned}
E(w) &= \frac{\beta}{2}||t-\Phi w||^2 + \frac{\alpha}{2}w^{\boldsymbol{T}}w \\
&= \frac{1}{2}w^{\boldsymbol{T}}(\alpha I+\beta\Phi^{\boldsymbol{T}}\Phi)w - \beta t^{\boldsymbol{T}}\Phi w + \frac{\beta}{2}||t||^2 \\
&= \frac{1}{2}w^{\boldsymbol{T}}Aw - \beta w^{\boldsymbol{T}}\Phi^{\boldsymbol{T}}t + \frac{\beta}{2}||t||^2.
\end{aligned}$$

ここで一般に対称行列 A とベクトル w, m について

$$\frac{1}{2}(w-m)^{\boldsymbol{T}}A(w-m) = \frac{1}{2}w^{\boldsymbol{T}}Aw - w^{\boldsymbol{T}}Am + \frac{1}{2}m^{\boldsymbol{T}}Am.$$

この関数は $w=m$ のとき最小値 0 をとる. 二つを比較することで $E(w)$ は $\beta\Phi^{\boldsymbol{T}}t=Am$, つまり

$$w=m_N=\beta A^{-1}\Phi^{\boldsymbol{T}}t$$

のとき最小となる. 最小値は元の $E(w)$ の式に $w=m_N$ を代入すれば得られ,

$$E(m_N) = \frac{\beta}{2}||t-\Phi m_N||^2 + \frac{\alpha}{2}m_N^{\boldsymbol{T}}m_N.$$

つまり

$$E(w) = \frac{1}{2}(w-m_N)^{\boldsymbol{T}}A(w-m_N) + E(m_N)$$

と平方完成できる.

よって

$$\begin{aligned}
E(w) &= \int \exp(-E(w))\,dw \\
&= \exp(-E(m_N))\int \exp(-\frac{1}{2}(w-m_N)^{\boldsymbol{T}}A(w-m_N))\,dw \\
&= \exp(-E(m_N))(2\pi)^{M/2}|A|^{-1/2}.
\end{aligned}$$

従って

$$\begin{aligned}
\log p(\boldsymbol{t}|\alpha,\beta) &= (N/2)\log(\frac{\beta}{2\pi}) + (M/2)\log(\frac{\alpha}{2\pi})\log(\int \exp(-E(w))\,dw) \quad (2) \\
&= (M/2)\log\alpha + \frac{N}{2}\log\beta - E(m_N) - \frac{1}{2}\log|A| - \frac{N}{2}\log(2\pi). \quad (3)
\end{aligned}$$

2.10　ヘッセ行列

x が n 次縦ベクトルのとき, $y=f(x)$ における 2 階微分の n 次正方行列

$$H(f)=(\frac{\partial^2}{\partial x_i x_j}f(x))$$

をヘッセ行列という．通常偏微分は可換なので, これは対称行列である．

1 階微分の行列（ヤコビ行列）の行列式はその点の付近の拡大率を表していた．ヘッセ行列はその点の付近の関数の形を表す．たとえば正定値な場合は極小, 固有値が全て負の場合は極大, 固有値が正と負の両方の場合は鞍点となる．

$f=x^2-y^2$, $g=x^2+y^2$ というグラフを見てみよう．図 1 は原点で鞍点, 図 2 は原点で極小である．それぞれヘッセ行列は

$$H(f)=\begin{pmatrix}2 & 0\\ 0 & -2\end{pmatrix},$$

$$H(g)=\begin{pmatrix}2 & 0\\ 0 & 2\end{pmatrix}$$

となり, ヘッセ行列が原点での形に対応していることが分かる．

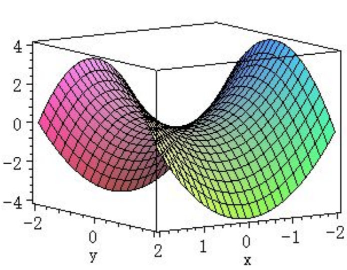

図 1　$f=x^2-y^2$

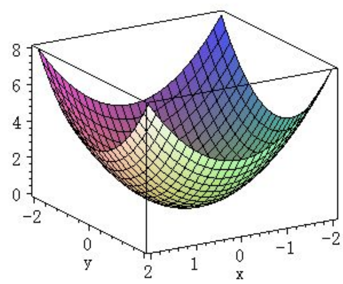

図 2　$g=x^2+y^2$

2.11　エビデンス関数の最大化の式変形

行列 $\beta\Phi^{\boldsymbol{T}}\Phi$ をある行列 P で対角化する．

$$P^{-1}(\beta\Phi^{\boldsymbol{T}}\Phi)P=\mathrm{diag}(\lambda_1,\dots,\lambda_M).$$

すると行列 $A=\alpha I+\beta\Phi^{\boldsymbol{T}}\Phi$ も同じ P で対角化できて

$$P^{-1}AP=\mathrm{diag}(\alpha+\lambda_1,\dots,\alpha+\lambda_M).$$

よって

$$|A|=\prod_{i=1}^{M}(\lambda_i+\alpha)$$

となる．α で微分すると

$$\frac{\partial}{\partial\alpha}\log|A| = \sum_{i=1}^{M}\frac{1}{\lambda_i+\alpha}.$$

式 2 を α で微分すると

$$\frac{\partial}{\partial\alpha}\log p(\boldsymbol{t}|\alpha,\beta) = \frac{M}{2\alpha} - \frac{1}{2}m_N^{\boldsymbol{T}}m_N - \frac{1}{2}\sum\frac{1}{\lambda_i+\alpha} = 0.$$

よって

$$\alpha m_N^{\boldsymbol{T}}m_N = M - \sum_{i=1}^{M}\frac{\alpha}{\lambda_i+\alpha} = \sum_{i=1}^{M}\frac{\lambda_i}{\lambda_i+\alpha}.$$

これを γ とおくと

$$\alpha = \frac{\gamma}{m_N^{\boldsymbol{T}}m_N}.$$

ただし, m_N は陰に α に依存しているのでこれは実は α を含む方程式である．

β についても同様にしてみる．$\beta\Phi^{\boldsymbol{T}}\Phi$ の固有値が λ_i だから λ_i は β に比例する．つまり微分が比例係数に等しい．

$$\frac{\partial}{\partial\beta}\lambda_i = \lambda_i/\beta.$$

よって

$$\frac{\partial}{\partial\beta}\log|A| = \sum\frac{\lambda_i/\beta}{\lambda_i+\alpha} = \frac{\gamma}{\beta}.$$

式 2 を β で微分すると

$$\frac{N}{2\beta} - \frac{1}{2}||\boldsymbol{t}-\Phi m_N||^2 - \frac{\gamma}{2\beta} = 0.$$

よって

$$\frac{1}{\beta} = \frac{1}{N-\gamma}||\boldsymbol{t}-\Phi m_N||^2.$$

2.12 パラメータの関係

パラメータがたくさんでてきたのでそれらの関係を見直してみよう．まず線形基底モデルを考えた．$\phi(x)$ を M 個の基底関数からなるベクトルとする．x は観測値であり,

$$y(x,w) = w^{\boldsymbol{T}}\phi(x)$$

とした．$\boldsymbol{t}$ を観測値に対する目標値で, それは x によらずに精度パラメータ β に従うガウス分布とした．

$$p(\boldsymbol{t}|w,\beta) = N(t|y(x,w),\beta^{-1}).$$

ベイズ的に扱うために w に関して事前確率分布を与えたいのだが, 上式が w に関する 2 次関数なので, 共役事前分布としてハイパーバラメータパラメータ α を導入し,

$$p(w|\alpha) = N(w|0,\alpha^{-1}I)$$

を仮定した．そうすることで事後分布は

$$p(w|t) = N(w|m_N,S_N)$$

の形（ただし, $m_N = \beta S_N \Phi^{\boldsymbol{T}} t$, $S_N^{-1} = \alpha I + \beta \Phi^{\boldsymbol{T}} \Phi$）になった.

さて, ここで α, β はハイパーパラメータではあるが, 事前分布を入れて確率変数的に扱いたい. その上で最尤推定の手法を用いて実際のデータから値を決めるという枠組みを経験ベイズという. そのとき t の予測分布は

$$p(t|\boldsymbol{t}) = \int p(t|w,\beta) p(w|\boldsymbol{t},\alpha,\beta) p(\alpha,\beta|\boldsymbol{t})\, dw d\alpha d\beta$$

となる. とはいえ, そのまま扱うのは難しいのでまずデータが十分たくさんあるとき, α, β は殆ど固定値, つまり α,β の分布はある特定の値 $\hat{\alpha}$, $\hat{\beta}$ にデルタ関数的に近づくと仮定しよう.

$$p(\alpha,\beta|\boldsymbol{t}) \sim \delta_{\alpha,\hat{\alpha}} \delta_{\beta,\hat{\beta}}.$$

そうすると

$$p(t|\boldsymbol{t}) \sim \int p(t|w,\hat{\beta}) p(w|\boldsymbol{t},\hat{\alpha},\hat{\beta})\, dw$$

となり予測分布は $\hat{\alpha}$, $\hat{\beta}$ を求めればよいということになる.

次に α, β を求める方法を考える. ベイズの定理から

$$p(\alpha,\beta|\boldsymbol{t}) \propto p(\boldsymbol{t}|\alpha,\beta) p(\alpha,\beta)$$

となる. ここで $p(\alpha,\beta)$ はほぼ平坦, つまり α, β の値によって変動しないという仮定を置く.

そうすると事後分布を最大化する α, β を求める最尤推定の問題は, 尤度関数を最大化する問題に近似できる. この尤度関数をエビデンスといい, この手法をエビデンス近似という. そして, $p(\boldsymbol{t}|\alpha,\beta)$ を最大化するための α, β の関係式を求めたのが前節であった.

以上のパラメータの関係を図 3 に示した. 実際には, 初期値 α, β を適当に決め, この図に従って計算して新しい α, β を求めたあと再度繰り返す. それが収束すればその値を採用する. ここではその収束性については議論しない.

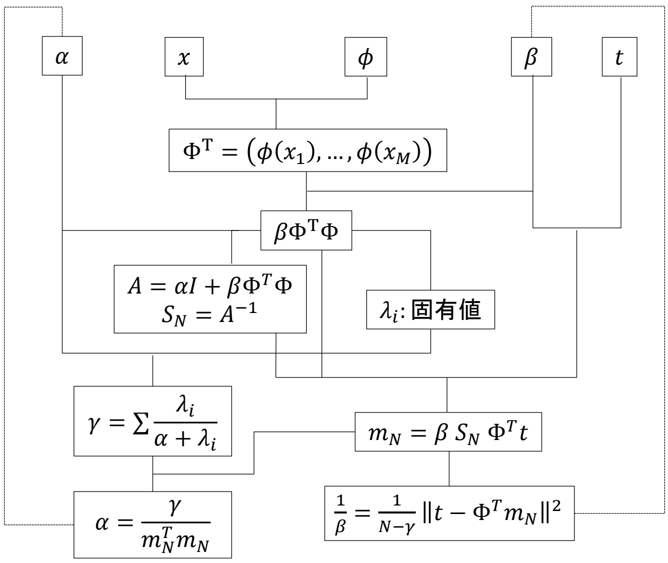

図 3　α,x,ϕ,t,β の関係図