PRML の 2 章のための数学

サイボウズ・ラボ 光成滋生

2011 年 5 月 18 日

1 概要

この文章は『パターン認識と機械学習』（以下 PRML）の 2 章, ガウス分布のところを理解するために必要な数学をまとめてみたものです. 目標はガウス分布の最尤推定の式変形をきちんと追えるようになることです.

いくつかの定理は証明せずに認めますが, 可能な限り self-contained であることを目指してみました. 概ね PRML に従ってますが, 違う方法をとっているところもあります.

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 微積分

2.1 変数変換

$$\int f(x)\,dx$$

で $x = g(y)$ とすると $dx = g'(y)dy$ より

$$\int f(g(y))g'(y)\,dy.$$

多変数関数の場合は $g'(y)$ の部分がヤコビ行列の行列式（ヤコビアン）になる.

$x_i = g(y_1, \ldots, y_n)$ for $i = 1, \ldots, n$ とすると

$$\det\left(\frac{\partial(x_1, \ldots, x_n)}{\partial(y_1, \ldots, y_n)}\right) = \det(\frac{\partial x_i}{\partial y_j}).$$

ヤコビアンは変数変換したときのある点における微小区間の拡大率を意味する.

2.2 奇関数の積分

全ての x について $f(-x) = -f(x)$ のとき f を奇関数という. $f(-x) = f(x)$ のときは偶関数という.

奇関数 f について

$$I = \int_{-\infty}^{\infty} f(x)\,dx = 0.$$

なぜなら $I = \int_{-\infty}^{0} f(x)\,dx + \int_{0}^{\infty} f(x)\,dx$ と積分区間を半分にわけてみよう. 第 1 項で $x = -y$ と変数変換すると $f(x)dx = -f(-y)dy = f(y)dy$ となる. 積分範囲は ∞ から 0 になり, 向きが逆転するので入れ換えると符号がひっくり返る. よって 第 1 項 $= -\int_{0}^{\infty} f(y)\,dy$. 第 2 項と打ち消しあって $I = 0$ となるからである.

$\boldsymbol{x}$ が n 次元ベクトルのときも同様に全ての $\boldsymbol{x}$ について $f(-\boldsymbol{x}) = -f(\boldsymbol{x})$ となるとき f を奇関数という.

やはり $I = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(\boldsymbol{x})\,dx_1 \cdots dx_n = 0.$

なぜなら

$$I = \int_{-\infty}^{0}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} + \int_{0}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}$$

と二つの領域に分けて $\boldsymbol{x} = -\boldsymbol{y}$ と変数変換すると, $d\boldsymbol{x} = (-1)^n d\boldsymbol{y}$.

第 1 項の積分範囲は $(0, -\infty) \times (\infty, -\infty) \times \cdots \times (\infty, -\infty)$ になり, 第 2 項の積分範囲に合わせると $(-1)^n$ がでる.

よって $f(-\boldsymbol{y}) = -f(\boldsymbol{y})$ を使うと

$$I = \int_{0}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(-\boldsymbol{y})\,d\boldsymbol{y} + \int_{0}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} f(\boldsymbol{x})\,d\boldsymbol{x} = 0.$$

2.3 $\exp(-x^2)$ の積分

$$I = \int_{0}^{\infty} \exp(-x^2)\,dx$$

とおくと

$$I^2 = \int_{0}^{\infty}\int_{0}^{\infty} \exp(-(x^2+y^2))\,dxdy.$$

ここで $x = r\cos(\theta)$, $y = r\sin(\theta)$ と置くと $x^2 + y^2 = r^2$.

ヤコビアンは

$$\det(\frac{\partial(x,y)}{\partial(r,\theta)}) = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r(\cos^2\theta + \sin^2\theta) = r.$$

積分範囲は x, y が (x, y) 平面の第一象限全体なので r は 0 から ∞, θ は 0 から $\pi/2$ を渡る. よって

$$I^2 = \int_{0}^{\pi/2}\int_{0}^{\infty} \exp(-r^2) r\,drd\theta = \pi/2\left[-\frac{1}{2}\exp(-r^2)\right]_{0}^{\infty} = \pi/4.$$

よって $I = \sqrt{\pi}/2$. x^2 は偶関数なので積分範囲を $-\infty$ から ∞ にすると 2 倍になって

$$\int_{-\infty}^{\infty} \exp(-x^2)\,dx = \sqrt{\pi}.$$

本当は積分の順序を交換したりしているところを気にしないといけないが, ここでは自由に交換できるものと思っておく.

2.4 ガウス分布の積分

前章の積分で $a > 0$ をとり $x = \sqrt{a}y$ とすると $dx = \sqrt{a}dy$.

$$\int_{-\infty}^{\infty} \exp(-x^2)\, dx = \int_{-\infty}^{\infty} \exp(-ay^2)\sqrt{a}\, dy = \sqrt{\pi}.$$

よって

$$\int_{-\infty}^{\infty} \exp(-ax^2)\, dx = \sqrt{\pi/a}.$$

ここで両辺を a に関して微分する. 積分の中身は $\frac{\partial}{\partial a}\exp(-ax^2) = -x^2\exp(-ax^2)$.
気にせず積分と微分を交換することで

$$-\int_{-\infty}^{\infty} x^2 \exp(-ax^2)\, dx = -1/2\sqrt{\pi}a^{-3/2}.$$

$a = 1/(2\sigma^2)$ と置き換えることで

$$\int_{-\infty}^{\infty} \exp(-\frac{1}{2\sigma^2}x^2)\, dx = \sqrt{2\pi}\sigma. \tag{1}$$

$$\int_{-\infty}^{\infty} x^2 \exp(-\frac{1}{2\sigma^2}x^2)\, dx = \sqrt{2\pi}\sigma^3. \tag{2}$$

式 (1) は正規化項が $\sqrt{2\pi}\sigma$ であることを示している. つまりガウス分布を

$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp(-\frac{1}{2\sigma^2}(x-\mu)^2)$$

とすると

$$\int_{-\infty}^{\infty} \mathcal{N}(x|\mu,\sigma^2)\, dx = 1.$$

平均は

$$x\mathcal{N}(x|\mu,\sigma^2) = (x-\mu)\mathcal{N}(x|\mu,\sigma^2) + \mu\mathcal{N}(x|\mu,\sigma^2)$$

とわけると, 第 1 項は $(x-\mu)$ に関して奇関数なので積分すると消えて

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} x\mathcal{N}(x|\mu,\sigma^2)\, dx = \mu.$$

分散は $x^2 = (x-\mu)^2 + 2\mu(x-\mu) + \mu^2$ なので積分すると第 1 項は式 (2) より σ^2. 第 2 項は 0. 第 3 項は μ^2.
よって

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} x^2\mathcal{N}(x|\mu,\sigma^2)\, dx = \sigma^2 + \mu^2.$$

3 線形代数

3.1 行列の積

以下, 特に断らない限り行列の数値は複素数とする.

A を m 行 n 列の行列とする. 横に n 個, 縦に m 個数字が並んでいる. A の i 行 j 列の値が a_{ij} であるとき, $A = (a_{ij})$ とかく. $m = n$ のとき n 次正方行列という. 並んでいる数字が実数値のみからなる行列を実行列という.

A を l 行 m 列の行列, B を m 行 n 列の行列とするとき, 積 AB を $(AB)_{ij} = \sum_{k=1}^{n} a_{ik}b_{kj}$ で定義する. AB は l 行 n 列の行列になる.

1. A, B が正方行列だったとしても $AB = BA$ とは限らない.
2. A, B, C がその順序で掛け算できるとき $(AB)C = A(BC)$ が成り立つ. なぜなら $((AB)C)_{ij} = \sum_k (AB)_{ik}c_{kj} = \sum_k (\sum_l a_{il}b_{lk})c_{kj} = \sum_{k,l} a_{il}b_{lk}c_{kj}$. $A(BC)_{ij} = \sum_l a_{il}(BC)_{lj} = \sum_l a_{il}(\sum_k b_{lk}c_{kj}) = \sum_{k,l} a_{il}b_{lk}c_{kj}$ だから.

3.2 トレース

A が n 次正方行列のとき $\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii}$ と A のトレースと呼ぶ.

$$\mathrm{tr}(A + B) = \mathrm{tr}(A) + \mathrm{tr}(B).$$

$$\mathrm{tr}(AB) = \mathrm{tr}(BA).$$

なぜなら $\mathrm{tr}(AB) = \sum_i (AB)_{ii} = \sum_i (\sum_j a_{ij}b_{ji}) = \sum_j (\sum_i b_{ji}a_{ij}) = \sum_j (BA)_{jj} = \mathrm{tr}(BA)$.

$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ のときは $\mathrm{tr}(A) = a + d$.

3.3 行列式

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

のとき, $|A| = ad - bc$ を A の行列式という. 一般には次のように定義する:

S_n を $1, \ldots, n$ の順序を並び替える操作全体の集合とする. たとえば S_2 は何も動かさない操作と 1 を 2 に, 2 を 1 に並び替える操作の二つの操作からなる. n 個の要素を並び替える組み合わせは $n \times (n-1) \times \cdots \times 1 = n!$ 通りある.

$D = \prod_{i<j}(x_i - x_j)$ とし, $S_n \ni \sigma$ に対して $\sigma D = \prod_{i<j}(x_{\sigma(i)} - x_{\sigma(j)})$ とすると,D と σD は符号しか変わらない. $\sigma D = \mathrm{sgn}(\sigma)D$ で $\mathrm{sgn}(\sigma) \in \{1, -1\}$ を定義する.

$\{\sigma(1), \ldots, \sigma(n)\}$ を 2 個ずつ順序を入れ換えて $\{1, \ldots, n\}$ に並び替えられたとき, 偶数回でできたら $\mathrm{sgn}(\sigma) = 1$, 奇数回でできたら $\mathrm{sgn}(\sigma) = -1$ である. これを使って行列式を定義する.

A を n 次正方行列（n 行 n 列）とするとき,

$$\det(A) = |A| = \sum_{\sigma \in S_n} \mathrm{sgn}(\sigma) a_{1\sigma(1)} \cdots a_{n\sigma(n)}.$$

A が 2 次正方行列のときを見直してみる．S_2 は 2 個の要素しかもたなかった．一つは何も動かさない操作でそれに対して sgn は 1．もう一つは 1 と 2 を入れ換える操作で sgn は -1 となる．よって

$$|A| = a_{11}a_{22} - a_{12}a_{21}.$$

ここで二つの n 次正方行列 A, B に対して $|AB| = |A||B|$ が成り立つ．2 次のときのみ確認しておこう．

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, B = \begin{pmatrix} x & y \\ z & w \end{pmatrix}$$

とするとき,

$$|AB| = \begin{pmatrix} ax+bz & ay+bw \\ cx+dz & cy+dw \end{pmatrix} = (ax+bz)(cy+dw) - (ay+bw)(cx+dz) = (ad-bc)(xw-yz) = \begin{vmatrix} a & b \\ c & d \end{vmatrix} \begin{vmatrix} x & y \\ z & w \end{vmatrix} = |A||B|.$$

一般のときの証明は省略する．

$|A|$ は負の数にもなることに注意する（絶対値の記号と間違えないように）．

3.4 行列の種類

A を m 行 n 列の行列とする．A に対して $A^{\boldsymbol{T}} = (a_{ji})$ を A の転置行列という．これは n 行 m 列の行列である．$|A^{\boldsymbol{T}}| = |A|$, $(AB)^{\boldsymbol{T}} = B^{\boldsymbol{T}}A^{\boldsymbol{T}}$ である．$\overline{A} = (\overline{a_{ij}})$ を A の複素共役行列, $A^* = \overline{A^{\boldsymbol{T}}}$ を随伴行列という．$|A^*| = \overline{|A|}$, $(AB)^* = B^*A^*$ である．

a_{ii} を対角成分という．対角成分以外の項が 0 である行列を対角行列といい $\mathrm{diag}(a_0, \ldots, a_n)$ と書く．

$$\delta_{ij} = \begin{cases} 1 & (i = j) \\ 0 & (i \neq j) \end{cases}$$

をクロネッカーの δ といい, $I_n = (\delta_{ij})$ を n 次単位行列という．I と略すこともあるし E と書くこともある．

A を n 次正方行列とすると $AI_n = I_nA = A$.

A が n 次正方行列で, $|A| \neq 0$ のとき A を正則といい, $AB = BA = I$ となる行列 B が存在する．B を逆行列といい, A^{-1} と書く．

1. 逆行列は存在すればただ一つである． なぜなら B, B' を逆行列とすると $B = BI = B(AB') = (BA)B' = IB' = B'$.
2. 有限次元では $AB = I$ ならば $BA = I$ であることを示すことができる．つまり $AB = I$ だが $BA \neq I$ なものは存在しない．

n 次正方行列 A について

1. $|A| \neq 0$ なもの全体を $GL_n(\mathbb{C})$ と書く．実正則行列全体は $GL_n(\mathbb{R})$ と書く．
2. $|A| = 1$ なものの全体を $SL_n(\mathbb{C})$ とかく．実行列のときは $SL_n(\mathbb{R})$.

3. $AA^* = I$ となるときユニタリー行列といい, その全体を $U(n)$ と書く. このとき $|AA^*| = ||A||^2 = 1$. ここで $||A||$ の内側の $||$ は行列式, 外側の $||$ は数値の絶対値である. ユニタリー行列であって, 更に $|A| = 1$ なもの全体を $SU(n)$ と書く.
4. 実行列 A が $AA^{\boldsymbol{T}} = I$ となるとき, 直交行列といい, その全体を $O(n)$ と書く. このとき $||A||^2 = 1$. 更に $|A| = 1$ なもの全体を $SO(n)$ と書く. $|A| \in \mathbb{R}$ なので $|A| = \pm 1$.
5. $A = A^{\boldsymbol{T}}$ となるとき対称行列という.

3.5 ブロック行列の逆行列

A, D を正方行列として（B, D は正方行列とは限らない）

$$X = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

の逆行列を求めてみよう.

逆行列を

$$X^{-1} = \begin{pmatrix} M & N \\ L & P \end{pmatrix}$$

とおくと

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} M & N \\ L & P \end{pmatrix} = \begin{pmatrix} AM + BL & AN + BP \\ CM + DL & CN + DP \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}.$$

2-1 ブロックに左から D^{-1} を掛けて $L = -D^{-1}CM$.

これを 1-1 ブロックに代入して $AM + BL = AM - BD^{-1}CM = (A - BD^{-1}C)M = I$.

よって $M = (A - BD^{-1}C)^{-1}$. 今度は

$$\begin{pmatrix} M & N \\ L & P \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} MA + NC & MB + ND \\ LA + PC & LB + PD \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix}$$

の 1-2 ブロックに右から D^{-1} を掛けて $N = -MBD^{-1}$.

2-2 ブロックに右から D^{-1} を掛けて $P = D^{-1} - LBD^{-1} = D^{-1} + D^{-1}CMBD^{-1}$.

よって $M = (A - BD^{-1}C)^{-1}$ として

$$\begin{pmatrix} M & N \\ L & P \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix}.$$

これが X の逆行列となることは容易に確認できる.

（以下余談）

$R = MB$, $S = D^{-1}C$ とおくと

$$X = \begin{pmatrix} M^{-1} & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I + MBD^{-1}C & MB \\ D^{-1}C & I \end{pmatrix} = \begin{pmatrix} M^{-1} & 0 \\ 0 & D \end{pmatrix} \begin{pmatrix} I + RS & R \\ S & I \end{pmatrix}$$

と変形できることはすぐ分かる.

$$\begin{pmatrix} M^{-1} & 0 \\ 0 & D \end{pmatrix}^{-1} = \begin{pmatrix} M & 0 \\ 0 & D^{-1} \end{pmatrix},$$

$$\begin{pmatrix} I+RS & R \\ S & I \end{pmatrix}^{-1} = \begin{pmatrix} I & -R \\ -S & I+SR \end{pmatrix}$$

なので X^{-1} もすぐ求められる. 更にこの行列の行列式は 1 なので

$$|X| = \begin{vmatrix} M^{-1} & 0 \\ 0 & D \end{vmatrix} = |M|^{-1}|D| = |A - BD^{-1}C||D|.$$

3.6 三角化

n 次正方行列 A に対して $a_{ij} = 0 (i > j)$ のとき（上半）三角行列という.

$$A = \begin{pmatrix} a_{11} & \dots & * \\ \vdots & \ddots & \vdots \\ 0 & \dots & a_{nn} \end{pmatrix}.$$

ここの $*$ は任意の値が入っていることを示す.

このとき $|A| = \prod_i a_{ii}$ である. なぜなら行列式の定義で 1 行ごとに異なる列のものをとっていったものの積を考えるわけだが, 最初に a_{11} 以外の $a_{1j} (j > 1)$ を選択すると, 残り $n-1$ 個をとる中で 0 でないものは $n-2$ 個しかない. したがって必ず 0 になる. 以下同様にして対角成分を拾ったものしか残らないからである.

さて次の定理を証明無しで認める:

任意の n 次正方行列 A に対して, あるユニタリー行列 P があって $P^{-1}AP$ を三角化できる. (3)

（注意）一般の行列が常に対角化できるとは限らないが三角化は常にできる.

3.7 対称行列

A を n 次実対称行列とする.

n 次実対称行列 A に対して, ある行列 P が存在して $P^{-1}AP$ を実対角化できる. (4)

定理 (3) を用いて証明しよう.

A に対してあるユニタリー行列 P があって $P^{-1}AP$ を三角化できる:

$$P^{-1}AP = \begin{pmatrix} \lambda_1 & \dots & * \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n \end{pmatrix}.$$

この両辺の随伴をとる. P はユニタリー行列なので $PP^* = I$. つまり $P^{-1} = P^*$. さらに A は実対称行列なので $A^* = A$ に注意すると

$$P^*A^*(P^{-1})^* = P^{-1}AP = \begin{pmatrix} \overline{\lambda_1} & \dots & 0 \\ \vdots & \ddots & \vdots \\ * & \dots & \overline{\lambda_n} \end{pmatrix}.$$

この二つの式が同一なので $\overline{\lambda_i} = \lambda_i$ かつ $*$ の部分が 0. これは $\lambda_i \in \mathbb{R}$ で, $P^{-1}AP$ はもともと対角行列であったことを意味する.

実は P が実行列であるようにもできる. そのときは P は直交行列になるので $|P| = \pm 1$.

もし $|P| = -1$ だったとすると, I' を単位行列の 1 行目と 2 行目を入れ換えたものとして $P' = PI'$ とおいて

$$P'^{-1}AP' = I'(P^{-1}AP)I' = I' \operatorname{diag}(\lambda_1, \ldots, \lambda_n)I' = \operatorname{diag}(\lambda_2, \lambda_1, \lambda_3, \ldots, \lambda_n).$$

これはもとの対角成分の 1 番目と 2 番目を入れ換えたものである. $|I'| = -1$, $|P'| = |P||I'| = 1$ なのでもともと $|P| = 1$ だったとしてもよい.

3.8 2 次形式

A を一般に n 次正方行列とし, $\boldsymbol{x}$ を n 次元縦ベクトルとする.

$$\boldsymbol{x}^{\boldsymbol{T}} A \boldsymbol{x} = \sum_i x_i (A\boldsymbol{x})_i = \sum_i x_i (\sum_j a_{ij} x_j) = \sum_{i,j} a_{ij} x_i x_j \tag{5}$$

を x の 2 次形式という.

A が与えられたときに $S = (A + A^{\boldsymbol{T}})/2$, $T = (A - A^{\boldsymbol{T}})/2$ とすると, $A = S + T$, $S^{\boldsymbol{T}} = S$, $T^{\boldsymbol{T}} = -T$ となる.

$T^{\boldsymbol{T}} = -T$ ということは $t_{ij} = -t_{ji}$ なので（標数 2 ではないから）$t_{ii} = 0$. 式 (5) の和を $i = j$ と $i \neq j$ の二つに分けて $A = T$ として適用すると

$$\boldsymbol{x}^{\boldsymbol{T}} T \boldsymbol{x} = \sum_i t_{ii} x_i x_j + \sum_{i<j} (t_{ij} + t_{ji}) x_i x_y.$$

第 1 項は $t_{ii} = 0$ より 0. 第 2 項も $t_{ij} = -t_{ji}$ より 0. つまり $T^{\boldsymbol{T}} = -T$ のとき 2 次形式の値は 0 となる.

よって $\boldsymbol{x}^{\boldsymbol{T}} A \boldsymbol{x} = \boldsymbol{x}^{\boldsymbol{T}} S \boldsymbol{x} + \boldsymbol{x}^{\boldsymbol{T}} T \boldsymbol{x} = \boldsymbol{x}^{\boldsymbol{T}} S \boldsymbol{x}$. つまり 2 次形式を考えるときは一般性を失うことなく A を対称行列としてよい.

2 変数のときを見てみる. 行列の計算は分かりにくければとりあえず 2 次で書いてみること.

$$\begin{pmatrix} x & y \end{pmatrix}^{\boldsymbol{T}} \begin{pmatrix} a & b \\ b & c \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} x & y \end{pmatrix}^{\boldsymbol{T}} \begin{pmatrix} ax + by \\ bx + cy \end{pmatrix} = ax^2 + 2bxy + cy^2.$$

ブロック行列なら A, C を対称行列として

$$\begin{pmatrix} \boldsymbol{x} & \boldsymbol{y} \end{pmatrix}^{\boldsymbol{T}} \begin{pmatrix} A & B \\ B^{\boldsymbol{T}} & C \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} = \begin{pmatrix} \boldsymbol{x} & \boldsymbol{y} \end{pmatrix}^{\boldsymbol{T}} \begin{pmatrix} A\boldsymbol{x} + B\boldsymbol{y} \\ B^{\boldsymbol{T}}\boldsymbol{x} + C\boldsymbol{y} \end{pmatrix} = \boldsymbol{x}^{\boldsymbol{T}} A \boldsymbol{x} + 2\boldsymbol{x}^{\boldsymbol{T}} B \boldsymbol{y} + \boldsymbol{y}^{\boldsymbol{T}} C \boldsymbol{y}.$$

ここで $\boldsymbol{x}^{\boldsymbol{T}} B \boldsymbol{y}$ はスカラー値なので転置しても変わらない, つまり

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

を用いた.

対称行列は $SO(n)$ の元 P を用いて対角化できた（$PP^{\boldsymbol{T}} = I$）. $y = P^{-1}x$ とおくと

$$\boldsymbol{x}^{\boldsymbol{T}} A \boldsymbol{x} = \boldsymbol{y}^{\boldsymbol{T}} P^{\boldsymbol{T}} A P \boldsymbol{y} = \boldsymbol{y}^{\boldsymbol{T}} \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \boldsymbol{y} = \sum_{i=1}^{n} \lambda_i y_i^2.$$

つまり 2 次形式は対角化すれば単なる成分ごとの直和になる.

3.9 多変量ガウス分布

A を n 次実対称行列, $\boldsymbol{z}$ を n 次元縦ベクトルとしてする.

まず $f(\boldsymbol{z}) = \exp(-(1/2)\boldsymbol{z}^{\boldsymbol{T}} A^{-1}\boldsymbol{z})$ を考える.

これは $\boldsymbol{z}$ について偶関数である.

A を直交行列 P で対角化する. $P^{-1}AP = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ より, $P^{-1}A^{-1}P = \mathrm{diag}(\lambda_1^{-1}, \ldots, \lambda_n^{-1})$.

$\boldsymbol{y} = P^{-1}\boldsymbol{z}$ と置いて前節の変形を行うと

$$f(\boldsymbol{z}) = \exp(-\frac{1}{2}\sum_{i=1}^{n}\frac{z_i^2}{\lambda_i}) = \prod_{i=1}^{n}\exp(-\frac{z_i^2}{2\lambda_i}).$$

ここで区間 $(-\infty, \infty)$ の積分を考えるが, そうすると積分値が発散しないためには全ての $\lambda_i > 0$ である必要がある. 以下この条件を仮定する. このとき $|A| = \prod_i \lambda_i > 0$.

積分値は式 (1) より

$$\int f(\boldsymbol{z})\, d\boldsymbol{z} = \prod_{i=1}^{n}\sqrt{2\pi\lambda_i} = \sqrt{2\pi}^n\sqrt{|A|}.$$

よって

$$\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, A) = \frac{1}{\sqrt{2\pi}^n}\frac{1}{\sqrt{|A|}}\exp(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\boldsymbol{T}}A^{-1}(\boldsymbol{x}-\boldsymbol{\mu}))$$

とすると正規化されている. これが多変量版のガウス分布である.

平均を求めよう:

$$\boldsymbol{x}f(\boldsymbol{x}-\boldsymbol{\mu}) = (\boldsymbol{x}-\boldsymbol{\mu})f(\boldsymbol{x}-\boldsymbol{\mu}) + \boldsymbol{\mu}f(\boldsymbol{x}-\boldsymbol{\mu})$$

とすると第 1 項は $(\boldsymbol{x}-\boldsymbol{\mu})$ に関して奇関数なので積分すると消える. 後者は $\boldsymbol{\mu}$ が定数で外に出るので

$$\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}.$$

次に分散を考える. $\boldsymbol{x}-\boldsymbol{\mu} = P\boldsymbol{y}$ とおくと

$$\boldsymbol{x}\boldsymbol{x}^{\boldsymbol{T}} = (P\boldsymbol{y}+\boldsymbol{\mu})(P\boldsymbol{y}+\boldsymbol{\mu})^{\boldsymbol{T}} = P\boldsymbol{y}\boldsymbol{y}^{\boldsymbol{T}}P^{\boldsymbol{T}} + P\boldsymbol{y}\boldsymbol{\mu}^{\boldsymbol{T}} + \boldsymbol{\mu}\boldsymbol{y}^{\boldsymbol{T}}P^{\boldsymbol{T}} + \boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{T}}.$$

これに $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, A)$ をかけた値を積分するわけだが, 第 2 項, 第 3 項は奇関数になるので 0. 第 4 項は $\boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{T}}$ が定数に出る.

第 1 項を考えよう: $P = (\boldsymbol{p_1}, \cdots, \boldsymbol{p_n})$ とすると $(P\boldsymbol{y})_i = \sum_{j=1}^{n} p_{ij}y_j$ だから

$$P\boldsymbol{y} = \sum_{j=1}^{n} y_j\boldsymbol{p_j}.$$

よって 第 1 項 $\times f(\boldsymbol{x}-\boldsymbol{\mu})$ は

$$\sum_{i,j}\boldsymbol{p_i}\boldsymbol{p_j}^{\boldsymbol{T}}y_iy_j\prod_{k=1}^{n}\exp(-\frac{y_k^2}{2\lambda_k}).$$

積分すると $i \neq j$ のところでは $y_i\exp(-y_i^2/(2\lambda_i))$ が奇関数になるので 0.

$i=j$ のところでは $y_i^2\exp(-\frac{y_i^2}{2\lambda_i})$ から λ_i がで, それ以外では 1. よって

$$\sum_i \boldsymbol{p_i}\boldsymbol{p_i^T}\lambda_i = (\boldsymbol{p_1},\cdots,\boldsymbol{p_n})\,\mathrm{diag}(\lambda_1,\cdots,\lambda_n)\begin{pmatrix}\boldsymbol{p_1^T}\\ \vdots \\ \boldsymbol{p_n^T}\end{pmatrix} = P\,\mathrm{diag}(\lambda_1,\cdots,\lambda_n)P^{\boldsymbol{T}} = A.$$

第 4 項と合わせて, 結局

$$\mathbb{E}[\boldsymbol{x}\boldsymbol{x^T}] = \boldsymbol{\mu}\boldsymbol{\mu^T} + A.$$

3.10 行列の微分

ここではガウス分布の最尤推定で使ういくつかの公式を列挙する. A を n 次正方行列とする.

1. 2 次形式の別の表現 $\boldsymbol{x}$ を n 次縦ベクトルとするとき,

$$\boldsymbol{x^T}A\boldsymbol{x} = \sum_{i,j} a_{ij}x_ix_j = \sum_i(\sum_j a_{ij}(\boldsymbol{x}\boldsymbol{x^T})_{ji}) = \sum_i(A\boldsymbol{x}\boldsymbol{x^T})_{ii} = \mathrm{tr}(A\boldsymbol{x}\boldsymbol{x^T}). \tag{6}$$

   この式は A が対称行列でなくても成り立つことに注意する.

2. 内積の微分

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

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

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

   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. $\frac{\partial}{\partial \boldsymbol{x}}$ を ∇ と書くこともあるが PRML では場所によって縦ベクトル（2.228）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.

   証明は $\boldsymbol{x^T}\boldsymbol{y} = \sum_j x_jy_j$ なので

$$\frac{\partial}{\partial x_i}(\boldsymbol{x^T}\boldsymbol{y}) = \sum_j \delta_{ij}y_j = y_j.$$

$$\frac{\partial}{\partial y_i}(\boldsymbol{x^T}\boldsymbol{y}) = \sum_j x_j\delta_{ij} = x_j.$$

3. 2 次形式の微分

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x^T}A\boldsymbol{x}) = (A+A^{\boldsymbol{T}})\boldsymbol{x}. \tag{7}$$

   証明は

$$\begin{aligned}\frac{\partial}{\partial x_i}(\boldsymbol{x^T}A\boldsymbol{x}) &= \sum_{s,t} a_{st}\frac{\partial}{\partial x_i}(x_sx_t) = \sum_{s,t} a_{st}(\delta_{is}x_t + x_s\delta_{it}) = (\sum_t a_{it}x_t) + (\sum_s a_{si}x_s)\\ &= (A\boldsymbol{x})_i + (A^{\boldsymbol{T}}\boldsymbol{x})_i = ((A+A^{\boldsymbol{T}})\boldsymbol{x})_i.\end{aligned}$$

   特に A が対称行列のときは

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x^T}A\boldsymbol{x}) = 2A\boldsymbol{x}.$$

4. $AA^{-1} = I$ の両辺を x で微分すると

$$(\frac{\partial}{\partial x}A)A^{-1} + A\frac{\partial}{\partial x}(A^{-1}) = 0.$$

左から A^{-1} をかけることによって

$$\frac{\partial}{\partial x}(A^{-1}) = -A^{-1}(\frac{\partial}{\partial x}A)A^{-1}. \tag{8}$$

5. 行列式の対数の微分の公式 (1)

$|A| > 0$ となる行列に対して

$$\frac{\partial}{\partial x}\log|A| = \mathrm{tr}(A^{-1}\frac{\partial}{\partial x}A).$$

（証明）A を P で三角化する:

$$A = P^{-1}\begin{pmatrix}\lambda_1 & \dots & * \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n\end{pmatrix}P.$$

ここで計算を見やすくするために

$$\begin{pmatrix}\lambda_1 & \dots & * \\ \vdots & \ddots & \vdots \\ 0 & \dots & \lambda_n\end{pmatrix} = \mathrm{tri}(\lambda_i)$$

と略記する. すると上の式は

$$A = P^{-1}\,\mathrm{tri}(\lambda_i)P$$

と表記できる. 逆行列は

$$A^{-1} = P^{-1}\,\mathrm{tri}(\lambda_i)^{-1}P$$

となる.

さて $|A| = \prod\lambda_i$ なので証明すべき式の左辺は

$$\frac{\partial}{\partial x}(\sum\log(\lambda_i)) = \sum\frac{\lambda_i'}{\lambda_i}.$$

ここで $\frac{\partial}{\partial x}\lambda_i = \lambda_i'$ と略記した.

証明すべき右辺を考えよう.

$$\frac{\partial}{\partial x}A = A' = (P^{-1}\,\mathrm{tri}(\lambda_i)P)' = (P^{-1})'\,\mathrm{tri}(\lambda_i)P + P^{-1}\,\mathrm{tri}(\lambda_i')P + P^{-1}\,\mathrm{tri}(\lambda_i)P'.$$

第 1 項に式 (8) を使うと

$$(P^{-1})'\,\mathrm{tri}(\lambda_i)P = -P^{-1}P'P^{-1}\,\mathrm{tri}(\lambda_i)P$$

更に $\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B)$ を使うと

$$\begin{aligned}\mathrm{tr}(A^{-1}A') &= -\mathrm{tr}((P^{-1}\,\mathrm{tri}(\lambda_i)^{-1}P)P^{-1}P'P^{-1}\,\mathrm{tri}(\lambda_i)P) \\ &\quad + \mathrm{tr}((P^{-1}\,\mathrm{tri}(\lambda_i)^{-1}P)P^{-1}\,\mathrm{tri}(\lambda_i')P) \\ &\quad + \mathrm{tr}((P^{-1}\,\mathrm{tri}(\lambda_i)^{-1}P)P^{-1}\,\mathrm{tri}(\lambda_i)P') \\ &= -\mathrm{tr}(P^{-1}\,\mathrm{tri}(\lambda_i)^{-1}P'P^{-1}\,\mathrm{tri}(\lambda_i)P) \\ &\quad + \mathrm{tr}(P^{-1}\,\mathrm{tri}(\lambda_i)^{-1}\,\mathrm{tri}(\lambda_i')P) \\ &\quad + \mathrm{tr}(P^{-1}P')\end{aligned}$$

次に $\mathrm{tr}(AB) = \mathrm{tr}(BA)$ を使ってトレースの中の積の順序を入れ換えて, 行列と逆行列の積を消していくと

$$\begin{aligned}\mathrm{tr}(A^{-1}A') &= -\mathrm{tr}(P'P^{-1}\,\mathrm{tri}(\lambda_i)PP^{-1}\,\mathrm{tri}(\lambda_i^{-1})) \\ &\quad + \mathrm{tr}(\mathrm{tri}(\lambda_i)^{-1}\,\mathrm{tri}(\lambda_i')PP^{-1}) \\ &\quad + \mathrm{tr}(P^{-1}P') \\ &= -\mathrm{tr}(P'P^{-1}) + \mathrm{tr}(\mathrm{tri}(\lambda_i)^{-1}\,\mathrm{tri}(\lambda_i')) + \mathrm{tr}(P^{-1}P') \\ &= \mathrm{tr}(\mathrm{tri}(\lambda_i)^{-1}\,\mathrm{tri}(\lambda_i')).\end{aligned}$$

三角行列の逆行列はやはり三角行列であり, $*$ の部分はもとの行列の部分とは異なる何かわからない値になる. しかし対角成分はもとの対角成分の逆数が並ぶ.
つまり

$$\mathrm{tri}(\lambda_i)^{-1} = \mathrm{tri}(\lambda_i^{-1}).$$

よって

$$\mathrm{tr}(A^{-1}A') = \mathrm{tr}(\mathrm{tri}(\lambda_i^{-1}\lambda_i')) = \sum \frac{\lambda_i'}{\lambda_i}.$$

これで左辺 = 右辺が示された.

6. 行列式の対数の微分の公式 (2)
$|A| > 0$ となる行列に対して

$$\frac{\partial}{\partial A}\log|A| = (A^{-1})^{\boldsymbol{T}}. \tag{9}$$

ここで行列 A で微分するというのは各要素 a_{ij} で微分したものを, 行列に並べたものを意味する.
今示した対数の微分の公式 (1) より

$$\frac{\partial}{\partial a_{ij}}\log|A| = \mathrm{tr}(A^{-1}\frac{\partial}{\partial a_{ij}}A).$$

$\frac{\partial}{\partial a_{ij}}A$ は ij 成分のみが 1 でそれ以外は 0 の行列になる.
その行列を I_{ij} と書くと,

$$\mathrm{tr}(A^{-1}I_{ij}) = \sum_s (A^{-1}I_{ij})_{ss} = \sum_s (\sum_t (A^{-1})_{st}(I_{ij})_{ts}) = \sum_s (\sum_t (A^{-1})_{st}\delta_{it}\delta_{js}) = (A^{-1})_{ji}.$$

つまり $\log|A|$ を a_{ij} 成分で微分すると A^{-1} の ji 成分になることが分かったので証明完了.
実はこの式は三角化を使わなくても行列式の定義から直接示すことができる. 2 次正方行列で示してみよう. $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ とすると $|A| = ad - bc$. よって左辺は $\log|A|$ を a, b, c, d でそれぞれ微分して

$$\text{左辺} = \frac{1}{|A|}\begin{pmatrix} d & -c \\ -b & a \end{pmatrix} = \text{右辺}.$$

一般のときは $|A| = \sum_{\sigma\in S_n}\mathrm{sgn}(\sigma)a_{1\sigma(1)}\cdots a_{n\sigma(n)}$ なので

$$|A|(\text{左辺})_{ij} = \sum_{\sigma\in S_n}\mathrm{sgn}(\sigma)\frac{\partial}{\partial a_{ij}}(a_{1\sigma(1)}\cdots a_{n\sigma(n)}).$$

a_{ij} による微分を考えると, 掛け算の中に a_{ij} があれば（微分が 1 なので）それを取り除き, なければ 0 になってしまう. a_{ij} が現れるのは $j = \sigma(i)$ を固定する σ についてのみである. つまり行列 A から i 行 j 列を取り除いたものになる.

実はこの式は A の余因子行列 $\tilde{A}$ の余因子 $\tilde{A}_{ji}$ と呼ばれるもので,

$$A\tilde{A} = |A|I$$

となることが示される（というか順序が逆で, 普通は逆行列をこれで構成する）. つまり 左辺 $= (A^{-1})^{\boldsymbol{T}}$.

3.11 ガウス分布の最尤推定

多変量ガウス分布から, N 個の観測値 $\boldsymbol{X} = \{\boldsymbol{x_i}\}$ が独立に得られたときに, 対数尤度関数

$$\log p(\boldsymbol{X}|\boldsymbol{\mu}, A) = -\frac{Nn}{2}\log(2\pi) - \frac{N}{2}\log|A| - \frac{1}{2}\sum_{i=1}^{N}(\boldsymbol{x_i} - \boldsymbol{\mu})^{\boldsymbol{T}}A^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu})$$

を A についての最大化を求めてみる. ここでは PRML では式変形の途中で対称性を利用しているが, せっかくなので A の対称性を仮定せずに話を進める.

その前にまず A を固定したときの $\boldsymbol{\mu}$ に関する最尤推定の解を求めておこう. 式 (7) より

$$\frac{\partial}{\partial \boldsymbol{\mu}}\log p(\boldsymbol{X}|\boldsymbol{\mu}, A) = \frac{1}{2}\sum_{i=1}^{N}(A^{-1} + (A^{-1})^{\boldsymbol{T}})(\boldsymbol{x_i} - \boldsymbol{\mu}) = \frac{1}{2}(A^{-1} + (A^{-1})^{\boldsymbol{T}})((\sum_{i=1}^{N}\boldsymbol{x_i}) - N\boldsymbol{\mu}).$$

これが $\boldsymbol{0}$ なので

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{i}\boldsymbol{x_i}.$$

さて, 本題に戻る. 再び A が対称行列でないという仮定に注意して式を変形する.

$\boldsymbol{y_i} = \boldsymbol{x_i} - \boldsymbol{\mu}$ とおき

$$F(A) = -N\log|A| - \sum_{i}\boldsymbol{y_i^T}A^{-1}\boldsymbol{y_i} = -N\log|A| - \mathrm{tr}(A^{-1}\sum_{i}\boldsymbol{y_i}\boldsymbol{y_i^T})$$

とおく. 第 2 項の式変形には式 (6) を用いた.

$B = \sum_i \boldsymbol{y_i}\boldsymbol{y_i^T}$ と置いて A で微分しよう.

第 1 項は式 (9) を使って $-N(A^{-1})^{\boldsymbol{T}}$. 第 2 項を求めるには式 (8) を使って

$$\frac{\partial}{\partial a_{ij}}\mathrm{tr}(A^{-1}B) = \mathrm{tr}((\frac{\partial}{\partial a_{ij}}A^{-1})B) = -\mathrm{tr}(A^{-1}(\frac{\partial}{\partial a_{ij}}A)A^{-1}B) = -\mathrm{tr}((\frac{\partial}{\partial a_{ij}}A)A^{-1}BA^{-1}).$$

最後の式変形では $\mathrm{tr}(XY) = \mathrm{tr}(YX)$ を使った. $C = A^{-1}BA^{-1}$ とおく.

$$\mathrm{tr}((\frac{\partial}{\partial a_{ij}}A)C) = \sum_{s}((\frac{\partial}{\partial a_{ij}}A)C)_{ss} = \sum_{s}(\sum_{t}(\frac{\partial}{\partial a_{ij}}A)_{st}c_{ts}) = \sum_{s,t}\delta_{is}\delta_{jt}c_{ts} = c_{ji}.$$

つまり

$$\frac{\partial}{\partial A}\mathrm{tr}(A^{-1}B) = -C^{\boldsymbol{T}} = -(A^{-1}BA^{-1})^{\boldsymbol{T}}.$$

よって

$$\frac{\partial}{\partial A}F(A) = -N(A^{-1})^{\boldsymbol{T}} + (A^{-1}BA^{-1})^{\boldsymbol{T}}.$$

これが 0 になるような A が $F(A)$ の最大値を与える.

転置をとって

$$-NA^{-1} + A^{-1}BA^{-1} = 0.$$

$$A = \frac{1}{N}B = \frac{1}{N}\sum_i \boldsymbol{y_i}\boldsymbol{y_i^T} = \frac{1}{N}\sum_i (\boldsymbol{x_i} - \boldsymbol{\mu})(\boldsymbol{x_i} - \boldsymbol{\mu})^{\boldsymbol{T}}.$$

この A は明らかに対称行列である. つまり A に関する対称性を仮定せずに最尤解を求めると A が対称行列となることが分かった.

また, $\boldsymbol{\mu}$ について特に条件も無いので, 先に $\boldsymbol{\mu}$ に関して最尤推定による解 $\boldsymbol{\mu}_{\mathrm{ML}}$ を代入すれば $\boldsymbol{\mu}$ と A を同時に最大化したものの解となることが分かる.