机器学习日记

诗小丁

2022 年 08 月 16 日

前言

笔者最近在做深度学习在心室分割领域的研究，越往深探索越感觉自己不论是在 Python 语法上还是深度学习的算法框架以及优化器的选择上都存在知识上的漏洞。在认识到自己的不足后，笔者决定从语言和算法上着手弥补自己的短板。在参考学习机器学习白板推导[1, 2]、徐亦达教授机器学习课件[3] 等相关资料后，笔者整理了学习机器学习与深度学习的成果与心得，旨在从公式原理和应用上进行剖析（希望能坚持到最后）。

目前已确定更新的内容包括：高斯分布、线性回归、线性分类、PCA、SVM、EM、概率图、CNN。剩下内容因为笔者也在学习中，所以会更新的慢一点。那我们废话不多说，一起开始吧！

关键词

算法；机器学习；深度学习；

目录

第一章　高斯分布

前期准备知识

a. 数学期望的性质

1. 设 C 为常数，则 $E\left(C\right)=C$

2. 设 X 是随机变量，C 是常数，则 $E\left(CX\right)=CE\left(X\right)$

3. 设 X、Y 是任意两个随机变量，则 $E\left(X+Y\right)=E\left(X\right)+E\left(Y\right)$

4. 设 X、Y 是相互独立的随机变量，则 $E\left(XY\right)=E\left(X\right)E\left(Y\right)$

b. 方差的性质

1. 设 C 为常数，则 $D\left(C\right)=0$

2. 设 X 是随机变量，C 是常数，则 $D\left(CX\right)=C^2D\left(X\right),D\left(X+C\right)=D\left(X\right)$

3. 设 X、Y 是任意两个随机变量，则 $D\left(X\right)=E\left(X^2\right)-E\left(X\right)^2$

4. 设 X、Y 是相互独立的随机变量，则 $D\left(X+Y\right)=D\left(X\right)+D\left(Y\right)$

c. 数学期望不一定等于均值

前者指服从某一概率分布的随机变量全体值如 $X\sim(\mu,\sigma^2)$，后者指某次试验中所有样本的均值，如 $\mu_{MLE}=\frac{1}{n}\sum_{i=1}^{n}x_i$

d. 矩阵的性质

1. 对称矩阵：$A=A^T$

2. 正交矩阵：$A^T=A^{-1},AA^T=E$

3. 对角化：$Q^{-1}AQ=\Lambda$

4. 特征值分解：$AX=X\Lambda\Rightarrow A=X\Lambda X^{-1}$

5. 正定矩阵：$X^TAX>0$

6. 半正定矩阵：$X^TAX\geq 0$

1.1 极大似然估计

已知

$$X = \begin{pmatrix} x_1 & x_2 & \cdots & x_n \end{pmatrix}^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{pmatrix}_{n\times p}$$

其中，$\theta \sim N\left(\mathrm{M}, \Sigma\right), x_i \in \mathbb{R}^P, x_i \sim \left(\mathrm{M}, \Sigma\right)$ 且满足独立同分布

$$MLE\left(Maximum\ Likelihood\ Estimation\right): \boxed{\theta_{MLE} = \arg\max_{\theta} P\left(X|\theta\right)}$$

简化问题为一维数据且服从高斯分布，令$p=1, \theta \sim \left(\mu, \sigma^2\right)$，则

$$\begin{aligned} logP\left(X|\theta\right) &= log\prod_{i=1}^{n} P\left(x_i|\theta\right) \\ &= \sum_{i=1}^{n} logP\left(x_i|\theta\right) \\ &= \sum_{i=1}^{n} log\frac{1}{\sqrt{2\pi}\ \sigma}e^{-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}} \\ &= \sum_{i=1}^{n} log\frac{1}{\sqrt{2\pi}} + log\frac{1}{\sigma} - \frac{\left(x_i-\mu\right)^2}{2\sigma^2} \end{aligned} \tag{1.1}$$

由 (1) 式，则

$$\begin{aligned} \mu_{MLE} &= \arg\max_{\mu} logP\left(X|\theta\right) \\ &\approx \arg\max_{\mu} -\sum_{i=1}^{n} \frac{\left(x_i-\mu\right)^2}{2\sigma^2} \\ &\approx \arg\min_{\mu} \sum_{i=1}^{n} \left(x_i-\mu\right)^2 \end{aligned} \tag{1.2}$$

由 (2) 式，则

$$\frac{\partial}{\partial\mu}\sum_{i=1}^{n}\left(x_i-\mu\right)^2 = \sum_{i=1}^{n} 2\left(x_i-\mu\right)\left(-1\right) = 0 \tag{1.3}$$

由 (3) 式，解得，

$$\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

再由 (1) 式，则

$$\begin{aligned} \sigma_{MLE}^2 &= \arg\max_{\sigma} logP\left(X|\theta\right) \\ &\approx \arg\max_{\sigma} -\overbrace{\sum_{i=1}^{n}\left(log\sigma + \frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right)}^{\mathscr{L}} \end{aligned} \tag{1.4}$$

由 (4) 式，则

$$\begin{aligned} \frac{\partial \mathscr{L}}{\partial \sigma} &= -\sum_{i=1}^{n}\left(\frac{1}{\sigma} + \frac{\left(x_i-\mu\right)^2}{2}\left(-2\right)\sigma^{-3}\right) = 0 \\ &\Rightarrow -\sum_{i=1}^{n}\left(\sigma^2 - \left(x_i-\mu\right)^2\right) = 0 \end{aligned} \tag{1.5}$$

由 (5) 式，解得，

$$\sigma_{MLE}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_{MLE}\right)^2$$

1.2 有偏 VS 无偏

先说结论，MLE 估计的均值是无偏的，而方差是有偏的。判断估计值是否有偏，从 $E\left(\mu_{MLE}/\sigma_{MLE}\right) = \mu/\sigma$ 下手。下面来证明：

1.2.1 均值无偏证明

$$E\left(\mu_{MLE}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n}\left(x_i\right) \tag{1.6}$$

考虑到 $X \sim (\mu, \sigma^2)$，则

$$\frac{1}{n}\sum_{i=1}^{n} E\left(x_i\right) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu \tag{1.7}$$

综合 (1.6) (1.7), 即

$$E\left(\mu_{MLE}\right) = \mu$$

因此，MLE 对均值的估计是无偏的

1.2.2 方差有偏证明

$$E\left(\sigma_{MLE}^2\right) = E\left(\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_{MLE}\right)^2\right) \tag{1.8}$$

又因为

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_{MLE}\right)^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(x_i^2 - 2\mu_{MLE}x_i + \mu_{MLE}^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\mu_{MLE}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n}\mu_{MLE}^2
\end{aligned}$$

考虑到 $\mu_{MLE} = \frac{1}{n}\sum_{i=1}^{n} x_i$，则

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \mu_{MLE}\right)^2 &= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\mu_{MLE}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n}\mu_{MLE}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\mu_{MLE} * \mu_{MLE} + \mu_{MLE}^2 \\
&= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu_{MLE}^2
\end{aligned}$$

代入 (1.8) 中，则

$$
\begin{aligned}
E\left(\sigma_{MLE}^2\right) &= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu_{MLE}^2\right) \\
&= E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2 - \left(\mu_{MLE}^2 - \mu^2\right)\right) \\
&= \underbrace{E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2\right)}_{①} - \underbrace{E\left(\mu_{MLE}^2 - \mu^2\right)}_{②}
\end{aligned}
$$

针对 ①，因为 $X \sim (\mu, \sigma^2)$，即$\mu = E\left(x_i\right)$，化简为

$$
\begin{aligned}
E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2\right) &= \frac{1}{n}\sum_{i=1}^{n} E\left(x_i^2\right) - \mu^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(E\left(x_i^2\right) - E\left(x_i\right)^2\right)
\end{aligned}
$$

我们知道，方差的定义为 $D\left(x\right) = E\left(x^2\right) - E\left(x\right)^2$，则 ① 最终为

$$
E\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \mu^2\right) = \sigma
$$

针对 ②，因为 $E\left(\mu_{MLE}\right) = \mu$，化简为

$$
\begin{aligned}
E\left(\mu_{MLE}^2 - \mu^2\right) &= E\left(\mu_{MLE}^2\right) - \mu^2 \\
&= E\left(\mu_{MLE}^2\right) - E\left(\mu_{MLE}\right)^2 \\
&= D\left(\mu_{MLE}\right)
\end{aligned}
$$

从上面可以知道，现在问题转化为求解 $D\left(\mu_{MLLE}\right)$

$$
\begin{aligned}
D\left(\mu_{MLLE}\right) &= D\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} D\left(x_i\right) \\
&= \frac{1}{n^2} * n * \sigma \\
&= \frac{\sigma}{n}
\end{aligned}
$$

因此，② 式最终为

$$E\left(\mu_{MLE}^2-\mu^2\right)=\frac{\sigma}{n}$$

综合 ① ②,得

$$\begin{aligned} E\left(\sigma_{MLE}^2\right) &= \sigma-\frac{\sigma}{n} \\ &= \frac{n-1}{n}\sigma \end{aligned}$$

我们可以看出，

$$E\left(\sigma_{MLE}^2\right)\neq\sigma$$

因此，MLE 对方差的估计是有偏的，若想实现无偏

$$\hat{\sigma}=\frac{n}{n-1}\sigma_{MLE}=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\mu_{MLE}\right)^2$$

1.3 多维高斯分布

已知多维高斯分布

$$X\sim N\left(\mu,\Sigma\right)=\frac{1}{\left(2\pi\right)^{\frac{p}{2}}\left|\Sigma\right|^{\frac{1}{2}}}e^{\overbrace{\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}\left(x-\mu\right)\right)}^{quadratic\quad form}}$$

其中，$x_i\in\mathbb{R}^P$，具体来说

$$X=\begin{pmatrix} x_1\\ x_2\\ \vdots\\ x_n \end{pmatrix},\mu=\begin{pmatrix} \mu_1\\ \mu_2\\ \vdots\\ \mu_n \end{pmatrix},\Sigma=\begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1p}^2\\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2p}^2\\ \vdots & \vdots & \vdots & \vdots\\ \sigma_{p1}^2 & \sigma_{p2}^2 & \cdots & \sigma_{pp}^2 \end{pmatrix}$$

1.3.1 概率角度观察

我们可以看出，$(x-\mu)^T\Sigma^{-1}\left(x-\mu\right)$ 是马氏距离。特别地，当 Σ 为单位矩阵时，马氏距离等价为欧式距离。那么，马氏距离具体为多少呢？我们来推导一下：

由于 σ 一般为实数，所以 Σ 为实对称矩阵，即一定能特征值分解，即

$$
\begin{aligned}
\Sigma &= U\Lambda U^T \\
&= (u_1, u_2, \cdots, u_p) \begin{pmatrix} \lambda_1 & \cdots & 0 \\ 0 & \lambda_2 & \vdots \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_p \end{pmatrix} \begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_p^T \end{pmatrix} \\
&= (u_1\lambda_1, u_2\lambda_2, \cdots, u_p\lambda_p) \begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_p^T \end{pmatrix} \\
&= \sum_{i=1}^{p} u_i \lambda_i u_i^T
\end{aligned}
$$

其中，$U = (u_1, u_2, \cdots, u_p)$ 为正定矩阵，满足 $U^T = U^{-1}, UU^T = U^TU = I$ 的关系，$\Lambda = diag\left(\lambda_i\right)$ 因此，

$$
\begin{aligned}
\Sigma^{-1} &= \left(U\Sigma U^T\right)^{-1} \\
&= \left(U^T\right)^{-1} \Sigma^{-1} U^{-1} \\
&= U\Sigma^{-1}U^T \\
&= \sum_{i=1}^{p} u_i \frac{1}{\lambda_i} u_i^T
\end{aligned}
$$

因此，

$$
\begin{aligned}
\Delta &= (x-\mu)^T \Sigma^{-1} (x-\mu) \\
&= (x-\mu)^T \sum_{i=1}^{p} u_i \frac{1}{\lambda_i} u_i^T (x-\mu) \\
&= \sum_{i=1}^{p} (x-\mu)^T u_i \frac{1}{\lambda_i} u_i^T (x-\mu)
\end{aligned}
$$

令 $y_i = (x-\mu)^T u_i$, 则

$$\begin{aligned}\Delta &= \sum_{i=1}^{p} y_i \frac{1}{\lambda_i} y_i^T \\ &= \sum_{i=1}^{p} \frac{y_i^2}{\lambda_i}\end{aligned}$$

当 $p=2$ 时，相当于有两个变量在 u_i 方向投影，$\Delta = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2}$。我们可以看出，此时 Δ 为椭圆方程，该过程如图 1-1 所示。这也解释了很多深度学习的参数寻优示例要用椭圆表示的原因。

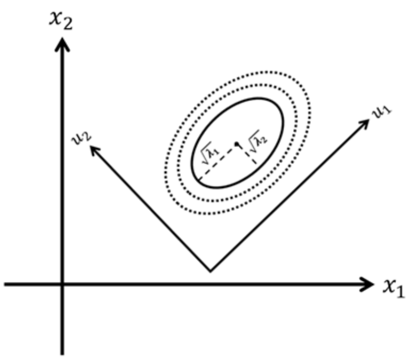

图 1-1 二维变量的概率切面

1.3.2 局限性

(1) 高斯分布的参数量取决于 μ 和 Σ 矩阵。经计算 Σ 矩阵的参数量为

$$O(\Sigma) = \frac{(1+p)\,p}{2} \sim (p^2)$$

相比于 p 维的 μ 矩阵，参数的计算量取决于 Σ 矩阵。随着 p 的增加，参数量逐渐增大，学习参数的计算量也逐渐增大，因此这是多维高斯分布的局限之一。那么如何解决参数爆炸的这一问题呢？

一个想法是，令 Σ 矩阵为对角矩阵，即

$$\Sigma = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_p \end{pmatrix}$$

此时 Σ 不需要做特征分解直接参与 Δ 的计算，u_i 变 x_i, 二维变量的概率在平面的切面如图 1-2 所示。特别地，当 $\lambda_1 = \lambda_2 = \cdots = \lambda_p$ 时，满足各项同性，此时 Δ 为圆的方程，二维变量的概率在平面的切面变为如图 1-3。

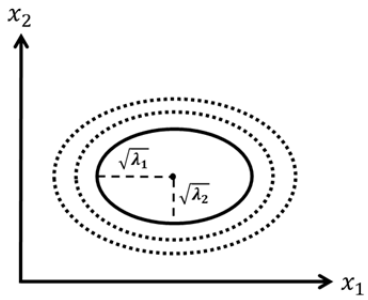

图 1-2 二维变量的概率切面 (对角矩阵)

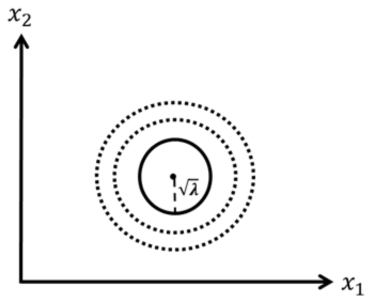

图 1-3 二维变量的概率切面（各向同性对角矩阵）

(2) 在有些情况下，单一的高斯分布不能代表分布的情况（有可能是混合高斯模型），这也是局限之一

1.3.3 求边缘概率和条件概率

已知

$$X = \begin{pmatrix} x_a \\ x_b \end{pmatrix} \begin{matrix} \Rightarrow & m \\ \Rightarrow & n \end{matrix}, \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix}, \Sigma = \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

求：

$$P(x_a), P(x_b|x_a), P(x_b), P(x_a|x_b)$$

定理

已知 $X \sim N(\mu, \Sigma), X \in \mathbb{R}^P$。$Y = AX + B, Y \in \mathbb{R}^P$, 则

$$X \sim N\left(A\mu + B, A\Sigma A^T\right)$$

证明：

$$\begin{aligned} E[Y] &= E[AX + B] \\ &= AE[X] + B \\ &= A\mu + B \end{aligned}$$

$$
\begin{aligned}
D[Y] &= D[AX+B] \\
&= AE[X]A^T \\
&= A\Sigma A^T
\end{aligned}
$$

(1) $P(x_a)$

我们首先将 x_a 分解成矩阵形式，即

$$
x_a = \underbrace{\begin{pmatrix} I_m & O_n \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_a \\ x_b \end{pmatrix}}_{X}
$$

根据定理，

$$
E[x_a] = \begin{pmatrix} I_m & O_n \end{pmatrix} \begin{pmatrix} x_a \\ x_b \end{pmatrix} = \mu_a
$$

$$
D[x_a] = \begin{pmatrix} I_m & O_n \end{pmatrix} \begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix} \begin{pmatrix} I_m \\ O_n \end{pmatrix} = \Sigma_{aa}
$$

因此，

$$
x_a \sim N(\mu_a, \Sigma_{aa}) \tag{1.9}
$$

(2) $P(x_b|x_a)$

根据构造法，令

$$
\begin{cases}
x_{b.a} &= x_b - \Sigma_{ba}\Sigma_{aa}^{-1}x_a \\
\mu_{b.a} &= \mu_b - \Sigma_{ba}\Sigma_{aa}^{-1}\mu_a \\
\Sigma_{bb.a} &= \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}
\end{cases}
$$

$$
x_{b.a} = \underbrace{\begin{pmatrix} -\Sigma_{ba}\Sigma_{aa}^{-1} & I_n \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} x_a \\ x_b \end{pmatrix}}_{X}
$$

根据构造内容，

$$
E[x_{b.a}] = \begin{pmatrix} -\Sigma_{ba}\Sigma_{aa}^{-1} & I_n \end{pmatrix} \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} = \mu_b - \Sigma_{ba}\Sigma_a^{-1}\mu_a
$$

$$D\left[x_{b.a}\right]=\left(\begin{array}{cc} -\Sigma_{ba}\Sigma_{aa}^{-1} & I_n \end{array}\right)\left(\begin{array}{cc} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{array}\right)\left(\begin{array}{c} -\Sigma_{ba}\Sigma_{aa}^{-1} \\ I_n \end{array}\right)=\Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}=\Sigma_{bb.a}$$

因此，

$$x_{b.a}\sim N\left(\mu_{b.a},\Sigma_{bb.a}\right) \tag{1.10}$$

我们为什么要先求 $x_{b.a}$ 呢？因为 $P\left(x_b|x_a\right)$ 与 $P\left(x_b\right)$ 有关，而 $P\left(x_b\right)$ 与 $P\left(x_{b.a}\right)$ 有关，所以我们必定要先求 $P\left(x_{b.a}\right)$。根据构造内容，即

$$\begin{aligned} x_b &= x_{b.a}+\Sigma_{b.a}\Sigma_{aa}^{-1}x_a \\ x_{b|a} &= x_{b.a|a}+\Sigma_{b.a}\Sigma_{aa}^{-1}x_{a|a} \end{aligned}$$

这就出现一个问题了,$x_{b.a}$ 与 x_a 独立吗？那我们引入一个定理来证明一下：

定理

若 $X\sim N\left(\mu,\Sigma\right)$ 则 $MX\perp NX\Leftrightarrow M\Sigma N^T=0$

证明：

高斯分布中独立【$\perp$】与不相关【协方差 $=0$】互为充要条件

$$\because X\sim N\left(\mu,\Sigma\right)$$

$$\begin{aligned} \therefore\quad & MX\sim N\left(M\mu,M\Sigma M^T\right) \\ & NX\sim N\left(N\mu,N\Sigma N^T\right) \end{aligned}$$

$$\begin{aligned} \therefore Cov\left(MX,NX\right) &= E\left[\left(MX-M\mu\right)\left(NX-N\mu\right)^T\right] \\ &= E\left[M\left(X-\mu\right)\left(X-\mu\right)^T N^T\right] \\ &= ME\left[\left(X-\mu\right)\left(X-\mu\right)^T\right]N^T \\ &= M\Sigma N^T \end{aligned}$$

(由左至右）

$$\because MX\perp NX \text{ 且均为高斯分布}$$
$$\therefore Cov\left(MX,NX\right)=M\Sigma N^T=0$$

(由右至左）

$$\because Cov\left(MX,NX\right)=M\Sigma N^{T}=0\text{ 且均为高斯分布}$$
$$\therefore MX\perp NX$$

$$x_a=\underbrace{\begin{pmatrix} I_m & O_n \end{pmatrix}}_{M}\underbrace{\begin{pmatrix} x_a \\ x_b \end{pmatrix}}_{X}$$

$$x_{b.a}=\underbrace{\begin{pmatrix} -\Sigma_{ba}\Sigma_{aa}^{-1} & I_n \end{pmatrix}}_{N}\underbrace{\begin{pmatrix} x_a \\ x_b \end{pmatrix}}_{X}$$

$$\Sigma=\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}$$

$$\begin{aligned}
\because M\Sigma N^T &= \begin{pmatrix} -\Sigma_{ba}\Sigma_{aa}^{-1} & I_n \end{pmatrix}\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\begin{pmatrix} I_m \\ O_n \end{pmatrix} \\
&= \begin{pmatrix} 0 & \Sigma_{bb}-\Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab} \end{pmatrix}\begin{pmatrix} I_m \\ O_n \end{pmatrix} \\
&= 0 \\
\therefore x_{b.a}\perp x_a &\Rightarrow x_{b.a}|x_a=x_{b.a}
\end{aligned}$$

因此 $x_{b.a}$ 与 x_a 相互独立，得证。那么，

$$\begin{aligned}
x_{b|a} &= x_{b.a|a}+\Sigma_{b.a}\Sigma_{aa}^{-1}x_{a|a} \\
&= x_{b.a}+\Sigma_{b.a}\Sigma_{aa}^{-1}x_{a|a}
\end{aligned}$$

因此，

$$E\left[x_{b|a}\right]=\mu_{b.a}+\Sigma_{b.a}\Sigma_{aa}^{-1}x_a$$

$$D\left[x_{b|a}\right]=D\left[x_{b.a}\right]=\Sigma_{bb.a}$$

$$x_{b|a} \sim N\left(\mu_{b.a} + \Sigma_{b.a}\Sigma_{aa}^{-1}x_a, \Sigma_{bb.a}\right) \tag{1.11}$$

同理

$$\begin{cases} x_{a.b} & = x_a - \Sigma_{ab}\Sigma_{bb}^{-1}x_b \\ \mu_{a.b} & = \mu_a - \Sigma_{ab}\Sigma_{bb}^{-1}\mu_b \\ \Sigma_{aa.b} & = \Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ab} \end{cases}$$

$$x_b \sim N\left(\mu_b, \Sigma_{bb}\right) \tag{1.12}$$

$$x_{a|b} \sim N\left(\mu_{a.b} + \Sigma_{a.b}\Sigma_{bb}^{-1}x_b, \Sigma_{aa.b}\right) \tag{1.13}$$

式 (1.9)、(1.11)、(1.12)、(1.13) 即为所求。

参考文献

[1] https://www.bilibili.com/video/BV1aE411o7qd

[2] https://www.yuque.com/books/share/f4031f65-70c1-4909-ba01-c47c31398466

[3] https://space.bilibili.com/327617676