Gaussian Distribution

1. MLE
2. PDF view
3. marginal Distri
4. Conditional Distri

Gaussian Distribution MLE

Data: $X = (x_1, x_2, \cdots, x_N)^T = \begin{pmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_N^T \end{pmatrix}_{N \times p}$

$x_i \in \mathbb{R}^p$ x_i iid $N(\mu, \Sigma)$. $\theta = (\mu, \Sigma)$

MLE: $\theta_{MLE} = \arg\max_{\theta} P(X|\theta)$

For $p=1$, $\theta = (\mu, \sigma^2)$

$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

$$\log P(X|\theta) = \log \prod_{i=1}^{N} P(x_i|\theta)$$
$$= \sum_{i=1}^{N} \log P(x_i|\theta)$$
$$= \sum_{i=1}^{N} \log \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$
$$= \sum_{i=1}^{N} \left[\log \frac{1}{\sqrt{2\pi}} + \log \frac{1}{\sigma} - \frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

$$\mu_{MLE} = \arg\max_{\mu} \log P(X|\theta)$$
$$= \arg\max_{\mu} \sum_{i=1}^{N} -\frac{(x_i-\mu)^2}{2\sigma^2} = \arg\min_{\mu} \sum_{i=1}^{N} (x_i-\mu)^2$$

$$\frac{d}{d\mu} \sum (x_i-\mu)^2 = \sum_{i=1}^{N} 2 \cdot (x_i-\mu) \cdot (-1) = 0$$

$$\sum_{i=1}^{N} (x_i-\mu) = 0$$

$$\boxed{\mu_{MLE} = \frac{1}{N} \sum_{i=1}^{N} x_i} \rightsquigarrow \text{无偏}$$

$$E[\mu_{MLE}] = \frac{1}{N} \sum_{i=1}^{N} E[x_i]$$
$$= \frac{1}{N} \cdot N \cdot \mu = \mu$$

$$\sigma^2_{MLE} = \arg\max_{\sigma} \log p(x|\theta)$$

$$= \arg\max_{\sigma} \sum_{i=1}^{N} \underbrace{\left[-\log\sigma - \frac{(x_i-\mu)^2}{2\sigma^2}\right]}_{L(\sigma)}$$

$$\frac{d}{d\sigma} L(\sigma) = \sum_{i=1}^{N} \left[-\frac{1}{\sigma} + (x_i-\mu)^2 \cdot \sigma^{-3}\right] = 0$$

So $\sum_{i=1}^{N} \left[-\sigma^2 + (x_i-\mu)^2\right] = 0$

$$N\sigma^2 = \sum_{i=1}^{N} (x_i-\mu)^2$$

So $$\boxed{\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu_{MLE})^2}$$

有偏估计

Because $E[\sigma^2_{MLE}] = \frac{N-1}{N}\sigma^2$

$$\sigma^2_{MLE} = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{MLE})^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - 2\mu_{MLE}x_i + \mu_{MLE}^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - 2\mu_{MLE}\cdot\frac{1}{N}\sum_{i=1}^{N} x_i + \frac{1}{N}\sum_{i=1}^{N}\mu_{MLE}^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - 2\mu_{MLE}^2 + \mu_{MLE}^2$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu_{MLE}^2$$

$$E[\sigma^2_{MLE}] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu_{MLE}^2\right]$$

$$= E\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right] - \left(\mu^2_{MLE} - \mu^2\right)$$

$$E\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right] = E\left[\frac{1}{N}\sum_{i=1}^{N}(x_i^2-\mu^2)\right] = \frac{1}{N}\sum_{i=1}^{N}\left[E(x_i^2-\mu^2)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[E(x_i^2) - E(\mu^2)\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N}\left[E(x_i^2) - \mu^2\right] = \frac{1}{N}\sum_{i=1}^{N}\underbrace{\mathrm{Var}(x_i)}_{\sigma^2}$$

$$= \sigma^2$$

$E[(\mu_{MLE}^2 - \mu^2)]$

$= E(\mu_{MLE}^2) - E(\mu^2)$

$= E(\mu_{MLE}^2) - \mu^2$

$= E(\mu_{MLE}^2) - E^2(\mu_{MLE})$

$Var(\mu_{MLE})$

$= \frac{1}{N}\sigma^2$

$Var(\mu_{MLE})$

$= Var\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right]$

$= \frac{1}{N^2} Var\left[\sum_{i=1}^{N} x_i\right]$

$= \frac{1}{N^2}\sum_{i=1}^{N} Var(x_i) = \frac{1}{N^2}\cdot\sum_{i=1}^{N}\sigma^2$

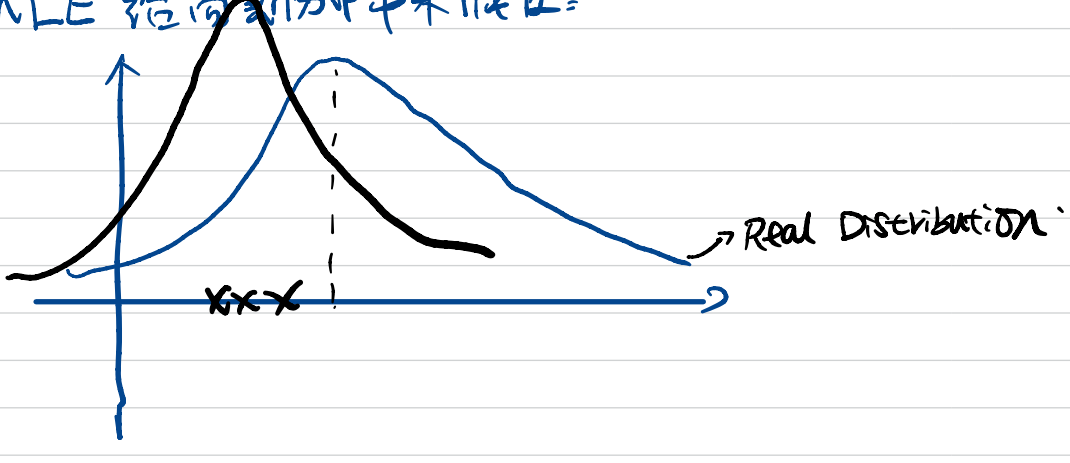
$= \frac{1}{N^2}\cdot N\cdot\sigma^2 = \frac{\sigma^2}{N}$

So $E[\sigma_{MLE}^2] = \frac{N-1}{N}\sigma^2 \neq \sigma^2$

σ_{MLE}^2 是 σ^2 的一个有偏估计。

$\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \mu_{MLE})^2$ → 无偏估计。

MLE 给高斯分布带来偏差：



从概率密度角度：(多维高斯)

PDF的值只与此次有关

$x \sim N(\mu, \Sigma)$，$x \in \mathbb{R}^p$，μ、Σ

此时 x 的 PDF：$\frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

$x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$ $\mu = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_2 \end{pmatrix}$ $\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots \\ \sigma_{21} & \sigma_{22} & \vdots \\ \cdots & \cdots & \end{pmatrix}_{p\times p}$ 实对称阵

Σ：正定的 → 半正定（一般）

$(x-\mu)^T\Sigma^{-1}(x-\mu) = D$

$1\times p \quad p\times p \quad p\times 1 \sim 1\times 1$

D为 x与μ的马氏距离

当 Σ 为 I 时，D为欧氏距离

$U = (u_1 \cdots u_p)_{p\times p}$

$\Sigma = U\Lambda U^T$，$UU^T = U^TU = I$，$\Lambda = \mathrm{diag}(\lambda_1 \cdots \lambda_p)$

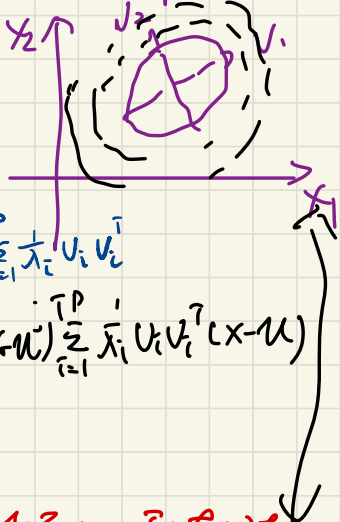
$\therefore \Sigma = (u_1, u_2 \cdots u_p)\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix}\begin{pmatrix} u_1^T \\ u_2^T \\ \vdots \\ u_p^T \end{pmatrix}$

$= (\lambda_1 u_1, \lambda_2 u_2 \cdots \lambda_p u_p)\begin{pmatrix} u_1^T \\ \vdots \\ u_p^T \end{pmatrix}$

$= \sum_{i=1}^{p} \lambda_i u_i u_i^T$

$\Sigma^{-1} = (U\Lambda U^T)^{-1} = (U^T)^{-1}\Lambda^{-1}U^{-1} = U\Lambda^{-1}U^T = \sum_{i=1}^{p}\frac{1}{\lambda_i}u_iu_i^T$

$\therefore (x-\mu)^T\Sigma^{-1}(x-\mu) = (x-\mu)^T\sum_{i=1}^{p}\frac{1}{\lambda_i}u_iu_i^T(x-\mu)$



故 $\Sigma^{-1} = \sum_{i=1}^{p}\frac{1}{\lambda_i}u_iu_i^T$

原式 $= \sum_{i=1}^{p}\frac{1}{\lambda_i}(x-\mu)^Tu_iu_i^T(x-\mu)$

令 $y = \begin{pmatrix} y_1 \\ \vdots \\ y_p \end{pmatrix}$，$y_i = (x-\mu)^Tu_i$ → 意义：x到中心在U轴的投影。

$= \sum_{i=1}^{p} y_i\frac{1}{\lambda_i}y_i^T = \sum_{i=1}^{p}\frac{y_i^2}{\lambda_i}$ → 由U轴上的椭圆

多x取不同值时，圆不同

形成等高线。

$\sum_{i=1}^{p}\frac{y_i^2}{\lambda_i}$ 就是一个以 y_i 为自变量的二次函数

形成等高线：

局限性:

① $\Sigma_{p\times p} \to \frac{p^2+p}{2}$

参数: $O(p^2)$

高维时计算太复杂

$\Sigma \to$ 对角阵 $\begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{pmatrix}$

↓ $\lambda_1=\lambda_2=\cdots=\lambda_p$

各向同性

例: factor analysis → Σ 对角矩阵

p-PCA → Σ 各向同性

② 一个高斯分布不足以描述数据data。

$$\frac{\partial L}{\partial \sigma} = \sum_{i=1}^{N} -\frac{1}{\sigma} + (x_i-\hat{\mu})^2 \sigma_{MLE}^{-3} = 0$$

$$\sum_{i=1}^{N} (x_i-\hat{\mu})^2 \sigma_{MLE}^{-3} = \frac{N}{\sigma_{MLE}}$$

$$\sigma_{MLE}^{-3} \sum_{i=1}^{N} (x_i-\hat{\mu})^2 = \frac{N}{\sigma_{MLE}}$$

$$\sum_{i=1}^{N} (x_i-\hat{\mu})^2 = N\sigma_{MLE}^2$$

$$\therefore \sigma_{MLE}^2 = \frac{\sum_{i=1}^{N}(x_i-\hat{\mu})^2}{N}$$ → 有偏估计

高斯—MLE(极大似然估计)

$$P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\hat{\theta} = \arg\max P(X|\theta)$$
$$= \arg\max \log P(X|\theta)$$
$$= \arg\max \log [P(x_1|\theta)P(x_2|\theta)\cdots P(x_N|\theta)]$$
$$= \arg\max \sum_{i=1}^{N} \log P(x_i|\theta)$$

$$\hat{\mu} = \arg\max_{\mu} \sum_{i=1}^{N} \log\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)\right)$$
$$= \arg\max_{\mu} \sum_{i=1}^{N}\left(\log\frac{1}{\sqrt{2\pi}} + \log\frac{1}{\sigma} - \frac{(x_i-\mu)^2}{2\sigma^2}\right)$$

↳ $L(\mu, \sigma^2)$

$$\frac{\partial L}{\partial \mu} = \sum_{i=1}^{N} -\frac{1}{2\sigma^2} 2(x_i-\mu)(-1)$$
$$= \sum_{i=1}^{N} \frac{x_i-\mu}{\sigma^2} = 0$$

$$N\mu = \sum_{i=1}^{N} x_i$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

↓ 无偏估计

$$\sigma_{MLE}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_{MLE})^2$$ 时无偏。

高维高斯分布求边缘概率分布概率

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

$X \in \mathbb{R}^p$, r.v.

$$X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix} \quad \mu = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_p \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & & & \\ \vdots & & & \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{pmatrix}_{p\times p}$$

已知：

$$x = \begin{pmatrix} x_a \to m \\ x_b \to n \end{pmatrix} \quad m+n=p \quad \mu = \begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} \quad \Sigma = \begin{pmatrix} \Sigma_{aa}\ (m\times m) & \Sigma_{ab}\ (m\times n) \\ \Sigma_{ba}\ (n\times m) & \Sigma_{bb}\ (n\times n) \end{pmatrix}$$

求 $P(X_a)$, $P(X_b|X_a)$ 两个分布 $\to$ PRML.

$P(X_a)$, $P(X_a|X_b)$

定理：已知 $X \sim N(\mu, \Sigma)$, $X \in \mathbb{R}^p$.

$Y = AX + B$

则 $Y \sim N(A\mu+B, A\Sigma A^T)$

So: $x_a = (\underbrace{I_m}_{A}, O_n)\begin{pmatrix} x_a \\ x_b \end{pmatrix}_{x}$ $(m\times 1)$

$$E[X_a] = (I_m \ O_n)\begin{pmatrix} \mu_a \\ \mu_b \end{pmatrix} = \mu_a$$

$$Var[X_a] = (I_m \ O_n)\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\begin{pmatrix} I_m \\ O_n \end{pmatrix} = \Sigma_{aa}$$

So $X_a \sim N(\mu_a, \Sigma_{aa})$

now for $x_b | x_a$.

we define a new variable $x_{b\cdot a}$.

$$x_{b\cdot a} = x_b - \Sigma_{ba}\Sigma_{aa}^{-1} x_a \quad (m\times 1)$$

$$= \underbrace{\left(-\underbrace{\Sigma_{ba}}_{n\times m}\underbrace{\Sigma_{aa}^{-1}}_{m\times m}, I_n\right)}_{A\ (n\times m)} \underbrace{\begin{pmatrix} x_a \\ x_b \end{pmatrix}}_{x\ (n\times 1)}$$

$$E[x_{b\cdot a}] = A\cdot\begin{pmatrix}\mu_a\\ \mu_b\end{pmatrix}$$

$$= \left(-\Sigma_{ba}\Sigma_{aa}^{-1}, I_n\right)\begin{pmatrix}\mu_a\\ \mu_b\end{pmatrix}$$

$$= -\Sigma_{ba}\Sigma_{aa}^{-1}\mu_a + \mu_b$$

$$Var[x_{b\cdot a}] = \left[-\Sigma_{ba}\Sigma_{aa}^{-1}, I_n\right]\begin{pmatrix}\Sigma_{aa} & \Sigma_{ab}\\ \Sigma_{ba} & \Sigma_{bb}\end{pmatrix}\begin{pmatrix}-\Sigma_{ba}\Sigma_{aa}^{-1}\\ I_n\end{pmatrix}$$

$$= \left(0 \quad \Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}\right)\begin{pmatrix}-\Sigma_{ba}\Sigma_{aa}^{-1}\\ I_n\end{pmatrix} = \underbrace{\Sigma_{bb} - \Sigma_{ba}\Sigma_{aa}^{-1}\Sigma_{ab}}_{\downarrow}$$

Σ_{aa} 的 Schur complement

So $x_{b\cdot a} \sim \mathcal{N}\left(-\Sigma_{ba}\Sigma_{aa}^{-1}\mu_a + \mu_b\right)$

Because

$$x_{b\cdot a} = x_b - \Sigma_{ba}\Sigma_{aa}^{-1} x_a$$

$$x_b = x_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1} x_a$$

$$x_b | x_a = \left(x_{b\cdot a} + \Sigma_{ba}\Sigma_{aa}^{-1} x_a\right) | x_a$$

Here we need prove $X_A \perp X_{S-A}$

Theorem: 对 $X \sim N(\mu, \Sigma)$, then $MX \perp NX \Leftrightarrow M\Sigma N^T = 0$

Prove:

$X \sim N(\mu, \Sigma)$

Then $MX \sim N(M\mu, M\Sigma M^T)$

$NX \sim N(N\mu, N\Sigma N^T)$

So $Cov(MX, NX)$

$= E[(MX - M\mu)(NX - N\mu)^T]$

$= E[M(X-\mu)\cdot(X-\mu)^T N]$

$= M \cdot E[(X-\mu)(X-\mu)^T] \cdot N$

$= M \cdot \Sigma \cdot N^T$

Because $MX \perp NX$ and they are both Gaussian 因为独立.

$\therefore Cov(MX, NX) = M\Sigma N^T = 0$

Note: ① 独立一定不相关，不相关不一定独立。

② 如果随个高斯分布，那么"不相关"等价于"独立".

Then: $X_{b\cdot a} = X_b - \bar{\Sigma}_{ba}\bar{\Sigma}_{aa}^{-1}X_a$.

$$= \underbrace{(-\bar{\Sigma}_{ba}\bar{\Sigma}_{aa}^{-1} \quad I_n)}_{M}\begin{pmatrix} X_a \\ X_b \end{pmatrix}_{X}$$

$$X_a = \underbrace{(I_n \quad 0)}_{N}\begin{pmatrix} X_a \\ X_b \end{pmatrix}_{X}.$$

$$\mathrm{Cov}(X_{b\cdot a}, X_a) = (-\bar{\Sigma}_{ba}\bar{\Sigma}_{aa}^{-1} \quad I_n)\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}\begin{pmatrix} I_n \\ 0 \end{pmatrix}$$

$$= 0$$

So $X_{b\cdot a} \perp X_a$

So $X_b | X_a = X_{b\cdot a} | X_a + \bar{\Sigma}_{ba}\bar{\Sigma}_{aa}^{-1} X_a | X_a$

$$= \underbrace{I}_{A}\underbrace{X_{b\cdot a}}_{x} + \underbrace{\bar{\Sigma}_{ba}\bar{\Sigma}_{aa}^{-1}X_a}_{B}$$

So

$$E[X_b | X_a] = \mu_{b\cdot a} + \bar{\Sigma}_{ba}\bar{\Sigma}_{aa}^{-1}X_a.$$

$$\mathrm{Var}(X_b | X_a) = \bar{\Sigma}_{bb\cdot a}$$

So

$$X_b | X_a \sim N(\mu_{b\cdot a} + \bar{\Sigma}_{ba}\bar{\Sigma}_{aa}^{-1}X_a, \bar{\Sigma}_{bb\cdot a})$$

已知: $p(x) = \mathcal{N}(x \mid \mu, \Lambda^{-1})$
$p(y \mid x) = \mathcal{N}(y \mid Ax+b, L^{-1})$

precision matrix = (covariance matrix)$^{-1}$

求 ① $p(y)$ · ② $p(x \mid y)$

解: Assume: $y = Ax + b + \varepsilon$, $x, y, \varepsilon \sim$ r.v., $\varepsilon \perp x$, $\varepsilon \sim \mathcal{N}(0, L^{-1})$

Then ① $E[y] = E[Ax+b+\varepsilon]$

$= E[Ax] + b$

$= A\mu + b$

$Var[y] = Var[Ax+b+\varepsilon]$

$= Var[Ax+\varepsilon]$

$= A\Lambda^{-1}A^T + L^{-1}$ ← from the theorem above.

So $y \sim \mathcal{N}(A\mu+b, L^{-1}+A\Lambda^{-1}A^T)$

② Let $z = \begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu \\ A\mu+b \end{bmatrix}, \begin{bmatrix} \Lambda^{-1} & \Delta \\ \Delta & L^{-1}+A\Lambda^{-1}A^T \end{bmatrix} \right)$

$\Delta = Cov(x, y)$

$= E[(x - Ex)(y - Ey)^T]$

$= E[(x-\mu)(y-A\mu-b)^T]$

$= E[(x-\mu)(Ax+b+\varepsilon-A\mu-b)^T]$

$$= E[(x-u)(Ax-Au+\varepsilon)^T]$$

$$= E[(x-u)(Ax-Au)^T] + E[(x-u)\varepsilon^T]$$

$$= E[(x-u)(Ax-Au)^T] + \underbrace{E[(x-u)]}_{0}\varepsilon^T$$

$$= E[(x-u)(Ax-Au)^T]$$

$$= E[(x-u)(x-u)^T A^T]$$

$$= \underbrace{E[(x-u)(x-u)^T]}_{\text{var}[x]} A^T$$

$$= \Lambda^{-1}A^T$$

Then based on

$$P(x_a|x_b) = N(\mu_{a\cdot b} + \Sigma_{ab}\Sigma_{bb}^{-1}x_b, \Sigma_{aa\cdot b})$$

$$E[x|y] = \mu + \Lambda^{-1}A^T(L^{-1} + A\Lambda^{-1}A^T)^{-1}(y - A\mu_b)$$

$$\text{var}[x|y] = \Lambda^{-1} - \Lambda^{-1}A^T(L^{-1} + A\Lambda^{-1}A^T)^{-1}A\Lambda^{-1}$$