2. Gaussian Distribution

(1) Introduction and Background

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

$$\mathcal{N}(\underline{x}|\underline{\mu},\underline{\Sigma})=\frac{1}{\sqrt{2\pi^D|\Sigma|}}\exp\left\{-\frac{1}{2}(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})\right\}$$

D is the dimension of $\underline{x}$

Because of very good properties for Gaussians, say Central Limit Theorem, we will focus greatly on this distribution.

the value

$$\Delta^2=(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})$$

is called Mahalanobis distance

Here we will state a very useful hypothesis that

> Matrix $\underline{\Sigma}$ can be viewed as symmetric

Actually, this assumption is not limited, We know that what really matters is the value $(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu})$, which is a scalar. Also, $(\underline{x}-\underline{\mu})^T(\underline{x}-\underline{\mu})$ is a scalar. So we can find a majic number λ such that $(\underline{x}-\underline{\mu})^T\lambda\cdot\underline{I}(\underline{x}-\underline{\mu})^T$. We find another matrix λI which could act the same effect as Σ^{-1}. So, why not choose a symmetric matrix?

Now, $\underline{\Sigma}$ is a real symmetric matrix. According to Spectrum Theorem, $\underline{\Sigma}$'s eigenvalues are real. $\underline{\Sigma}$ can be decomposed as

$$\underline{\Sigma} = \sum_{i=1}^{r} \lambda_i q_i q_i^T$$

r is the rank of $\underline{\Sigma}$. Here it must be full rank. So $r = D$

Then $$\underline{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} q_i q_i^T$$

So, $$\Delta^2 = (\underline{X} - \underline{\mu})^T \left(\sum \frac{1}{\lambda_i} q_i q_i^T \right) (\underline{X} - \underline{\mu})$$

$$= \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i} \qquad y_i = (\underline{X} - \underline{\mu})^T \cdot q_i$$

So if we can have a group of point's which could guarantee $y_i = (\underline{X} - \underline{\mu})^T q_i$ to be a constant, then the corresponding Gaussian distribution is constant. Mathematically, if all the λ_i are possitive, points $\underline{X}$ are on a ellipse. λ_i's the axis length. Actually, $\{\lambda_i\}$ are non-negative, or the Gaussian distribution will not be normalized. In other word, $\underline{\Sigma}$ is PSD.

(2) Conditional and Marginal Gaussian Distribution

First, we throw out the conclusion.

If two sets of variables are jointly Gaussian, the conditional distribution of one set on the other is also Gaussian. Marginal Distribution is again Gaussian

Now we will prove these important conclusions.

① Conditional Distribution $P(\underline{x}_a|\underline{x}_b)$

Now $\underline{x}$ is a D-dimentional vector with Gaussian distribution $\mathcal{N}(\underline{x}|\underline{\mu},\underline{\Sigma})$. Let's split it into two parts $\underline{x}_a$, $\underline{x}_b$.

$$\underline{x} = \begin{pmatrix} \underline{x}_a \\ \underline{x}_b \end{pmatrix} \begin{matrix}]M \\ ]D-M \end{matrix} \qquad \underline{\mu} = \begin{pmatrix} \underline{\mu}_a \\ \underline{\mu}_b \end{pmatrix} \qquad \underline{\Sigma} = \begin{pmatrix} \Sigma_a & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_b \end{pmatrix}$$

Note that Σ is symmetric, so $\Sigma_{ab} = \Sigma_{ba}^T$, Σ_a & Σ_b are symmetric.

Let's denote

$$\Lambda = \Sigma^{-1} \qquad \underline{\Lambda} = \begin{pmatrix} \Lambda_a & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_b \end{pmatrix}$$

Λ is also a symmetric matrix. $\Lambda^T = \Lambda$

The exponent part

$$-\frac{1}{2}(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu}) = -\frac{1}{2}\begin{pmatrix} \underline{x}_a-\underline{\mu}_a \\ \underline{x}_b-\underline{\mu}_b \end{pmatrix}^T \begin{pmatrix} \Lambda_a & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_b \end{pmatrix} \begin{pmatrix} \underline{x}_a-\underline{\mu}_a \\ \underline{x}_b-\underline{\mu}_b \end{pmatrix}$$

$$= -\frac{1}{2}\left\{ \begin{matrix} (\underline{x}_a-\underline{\mu}_a)^T\Lambda_a(\underline{x}_a-\underline{\mu}_a) + (\underline{x}_a-\underline{\mu}_a)^T\Lambda_{ab}(\underline{x}_b-\underline{\mu}_b) \\ + (\underline{x}_b-\underline{\mu}_b)^T\Lambda_{ba}(\underline{x}_a-\underline{\mu}_a) + (\underline{x}_b-\underline{\mu}_b)^T\Lambda_b(\underline{x}_b-\underline{\mu}_b) \end{matrix} \right\} \quad \text{---} \quad (*)$$

We again get quadratic form. because we get terms like $a^T \Sigma b$

So the result will also be a Gaussian distribution.

Our target now turns to find the mean $\mu_{a|b}$ and Covariance $\Sigma_{a|b}$.

Our method is coefficient analysis.

Because x_b is given, we can simply treat μ_b, x_b as constant.

For a quadratic form

$$-\frac{1}{2}(\underline{x}-\underline{\mu})^T \Sigma^{-1}(\underline{x}-\underline{\mu}) = -\frac{1}{2}\underline{x}^T\underline{\Sigma}^{-1}\underline{x} + \underline{x}^T\underline{\Sigma}\underline{\mu} - \underbrace{\frac{1}{2}\underline{\mu}^T\underline{\Sigma}\underline{\mu}}$$

We can find out the coefficient corresponding to $x^T x$ and x

no $\underline{x}$, constant !

In (X) expression, we have

$$-\frac{1}{2}\underline{x}_a^T \Lambda_a \underline{x}_a \backsim -\frac{1}{2}\underline{x}^T\underline{\Sigma}^{-1}\underline{x}$$

$$\underline{x}_a^T\{\Lambda_a \underline{\mu}_a - \Lambda_{ab}(\underline{x}_b - \underline{\mu}_b)\} \backsim \underline{x}^T\underline{\Sigma}\underline{\mu}$$

So, we have

$$\Lambda_{a|b} = \Lambda_a \,, \quad \mu_{a|b} = \Lambda_a^{-1}\{\Lambda_a \underline{\mu}_a - \Lambda_{ab}(\underline{x}_b - \underline{\mu}_b)\}$$
$$= \underline{\mu}_a - \Lambda_a^{-1}\Lambda_{ab}(\underline{x}_b - \underline{\mu}_b)$$

For partition matrix

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix}^{-1} = \begin{pmatrix} M & -MBD^{-1} \\ -D^{-1}CM & D^{-1} + D^{-1}CMBD^{-1} \end{pmatrix}$$

$M = (A - BD^{-1}C)^{-1}$

Schur complement r.p.t D

So $\Lambda_a = (\Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ba})^{-1}$

$\Lambda_{ab} = -(\Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}$

Now, the mean and variance for $\underline{x}_a|\underline{x}_b$ is

$$\underline{\mu}_{a|b} = \mu_a + \Sigma_{ab}\Sigma_b^{-1}(\underline{x}_b - \underline{\mu}_b)$$
$$\Sigma_{a|b} = \Sigma_a - \Sigma_{ab}\Sigma_b^{-1}\Sigma_{ba}$$

$\Rightarrow$ no $\underline{x}_a$, only linearly dependent on $\underline{x}_b$

---

② Marginal Gaussian Distribution

Marginal Distribution for $\underline{x}_a$ is

$$p(\underline{x}_a) = \int_{\underline{x}_b} p(\underline{x}_a, \underline{x}_b)\, d\underline{x}_b$$

So we will extract all terms involving $\underline{x}_b$ and treat the rest as constant.

According to (*) 2 pages before, terms relate to $\underline{x}_b$ are

$$-\frac{1}{2}\underline{x}_b^T \Lambda_b \underline{x}_b + \underline{x}_b^T \underline{m} \qquad \underline{m} = \Lambda_b \mu_b - \Lambda_{ba}(\underline{x}_a - \underline{\mu}_a)$$

make it a quadratic form

$$-\frac{1}{2}(\underline{x}_b - \Lambda_b^{-1}\underline{m})^T \Lambda_b (\underline{x}_b - \Lambda_b^{-1}\underline{m}) + \underbrace{\frac{1}{2}\underline{m}^T \Lambda^{-1} \underline{m}}_{\text{no } \underline{x}_b \text{ terms}}$$

$$\int \underbrace{\exp\left\{-\frac{1}{2}(\underline{x}_b - \Lambda_b^{-1}\underline{m})^T \Lambda_b (\underline{x}_b - \Lambda_b^{-1}\underline{m})\right\}}_{\text{un-normalized Gaussian}} d\underline{x}_b = \sqrt{(2\pi)^{D-M}|\Sigma_b|}$$

Now all the terms independent from $\underline{x}_b$ are

$$\frac{1}{2}m^T\Lambda_b^{-1}m - \frac{1}{2}\underline{x}_a^T\Lambda_a\underline{x}_a + \underline{x}_a^T(\Lambda_a\underline{\mu}_a + \Lambda_{ab}\underline{\mu}_b) + \text{const}$$

↳ no $\underline{x}_a, \underline{x}_b$

$$= \frac{1}{2}[\Lambda_b\mu_b - \Lambda_{ba}(\underline{x}_a - \underline{\mu}_a)]^T\Lambda_b^{-1}[\Lambda_b\mu_b - \Lambda_{ba}(\underline{x}_a - \underline{\mu}_a)]$$
$$- \frac{1}{2}\underline{x}_a^T\Lambda_a\underline{x}_a + \underline{x}_a^T(\Lambda_a\underline{\mu}_a + \Lambda_{ab}\underline{\mu}_b) + \text{const}$$

$$= -\frac{1}{2}\underline{x}_a^T\underbrace{(\Lambda_a - \Lambda_{ab}\Lambda_b^{-1}\Lambda_{ba})}_{\Sigma_a^{-1}}\underline{x}_a + \underline{x}_a^T(\Lambda_a - \Lambda_{ab}\Lambda_b^{-1}\Lambda_{ba})\underline{\mu}_a + \text{const}$$

$$= -\frac{1}{2}\underline{x}_a^T\Sigma_a^{-1}\underline{x}_a + \underline{x}_a^T\Sigma_a^{-1}\mu_a + \text{const}$$

Recall that, a valid quadratic form has.

$$-\frac{1}{2}(\underline{x}-\underline{\mu})^T\Sigma^{-1}(\underline{x}-\underline{\mu}) = -\frac{1}{2}\underline{x}^T\Sigma^{-1}\underline{x} + \underline{x}^T\Sigma^{-1}\underline{\mu} + \text{const}$$

So, here we have the mean and variance for marginal distribution of $\underline{x}_a$

$$\mathbb{E}(\underline{x}_a) = \underline{\mu}_a \quad \text{Cov}(\underline{x}_a) = \Sigma_a$$

(3) Bayes' Theorem for Gaussian Variables

In section (2), we give out the mean and variance for conditional and marginal Gaussian variables. Which is

$$\text{Given } \begin{matrix} P(\underline{x},\underline{y}) \\ \underline{\mu}_x \ \underline{\mu}_y \\ \Sigma_x, \Sigma_y, \Sigma_{xy}, \Sigma_{yx} \end{matrix} \Longrightarrow \begin{matrix} P(\underline{x}|\underline{y}) & P(\underline{y}|\underline{x}) \\ P(\underline{x}) & P(\underline{y}) \end{matrix}$$

In this section, we will apply Bayes Law to do the inverse. We know that, $P(\underline{x}|\underline{y})$ is a linear function of $\underline{y}$. So lets use these start condition

$$\text{Given } P(\underline{x}) = \mathcal{N}(\underline{x}\,|\,\underline{\mu}, \Lambda^{-1})$$

$$P(\underline{y}|\underline{x}) = \mathcal{N}(\underline{y}\,|\,A\underline{x}+\underline{b}, L^{-1})$$

① find out the joint distribution

$$\text{let } \underline{z} = \begin{pmatrix} \underline{x} \\ \underline{y} \end{pmatrix}$$

$$\begin{aligned} \log p(\underline{z}) &= \ln p(\underline{x}) + \ln p(\underline{y}|\underline{x}) \\ &= -\frac{1}{2}(\underline{x}-\underline{\mu})^T \Lambda (\underline{x}-\underline{\mu}) - \frac{1}{2}(\underline{y}-A\underline{x}-b)^T L (\underline{y}-A\underline{x}-\underline{b}) + \text{const} \end{aligned}$$

Consider all the second order term (term involves two x or two y or x and y).

$$-\frac{1}{2}\underline{x}^T(\Lambda + A^T L A)\underline{x} - \frac{1}{2}\underline{y}^T L \underline{y} + \frac{1}{2}\underline{y}^T L A \underline{x} + \frac{1}{2}\underline{x}^T A^T L \underline{y}$$

$$= -\frac{1}{2}\begin{pmatrix} \underline{x} \\ \underline{y} \end{pmatrix}^T \underbrace{\begin{pmatrix} \Lambda + A^T L A & -A^T L \\ -LA & L \end{pmatrix}}_{R} \begin{pmatrix} \underline{x} \\ \underline{y} \end{pmatrix} = -\frac{1}{2}\underline{z}^T R \underline{z}$$

Again comparing this with the expansion of quadratic form of Gaussian,

$$\mathrm{Cov}(\underline{z}) = R^{-1} = \begin{pmatrix} \Lambda^{-1} & \Lambda^{-1}A^T \\ A\Lambda^{-1} & L^{-1} + A\Lambda^{-1}A^T \end{pmatrix}$$

Now inspect the linear terms (involving only one x or y)

$$\underline{x}^T\Lambda\underline{\mu} - \underline{x}^T A^T L\underline{b} + \underline{y}^T L\underline{b} = \begin{pmatrix} \underline{x} \\ \underline{y} \end{pmatrix}^T \begin{pmatrix} \Lambda\underline{\mu} - A^T L\underline{b} \\ L\underline{b} \end{pmatrix}$$

So

$$E(\underline{z}) = R^{-1}\begin{pmatrix} \Lambda\underline{\mu} - A^T L\underline{b} \\ L\underline{b} \end{pmatrix} = \begin{pmatrix} \underline{\mu} \\ A\underline{\mu} + \underline{b} \end{pmatrix}$$

(4) Maximum Likelihood for Gaussian.

Now our dataset is $X = (\underline{x}_1, \underline{x}_2, \cdots, \underline{x}_N)^T$. The log likelihood is

$$\ln p(X|\underline{\mu}, \Sigma) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(\underline{x}_n - \underline{\mu})^T\Sigma^{-1}(\underline{x}_n - \underline{\mu})$$

$$\frac{\partial}{\partial \underline{\mu}}\ln p(X|\underline{\mu}, \Sigma) = \sum_{n=1}^{N}\Sigma^{-1}(\underline{x}_n - \underline{\mu}) = 0 \Rightarrow \boxed{\underline{\mu}_{ML} = \frac{1}{N}\sum_{n=1}^{N}\underline{x}_n}$$

$$\frac{\partial}{\partial \Sigma}\ln p(X|\underline{\mu}, \Sigma) = \frac{\partial}{\partial \Sigma} -\frac{N}{2}\ln|\Sigma| - \frac{1}{2}\sum_{n=1}^{N}(\underline{x}_n - \underline{\mu})^T\Sigma^{-1}(\underline{x}_n - \underline{\mu})$$

According to «The Matrix Cookbook»

$$\frac{\partial}{\partial \Sigma} -\frac{N}{2}\ln|\Sigma| = -\frac{N}{2}\Sigma^{-T}$$

$$\frac{\partial}{\partial \Sigma} -\frac{1}{2}\sum_{n=1}^{N}(\underline{x}_n - \underline{\mu})^T\Sigma^{-1}(\underline{x}_n - \underline{\mu}) = \frac{1}{2}\sum_{n=1}^{N}\Sigma^{-T}(\underline{x}_n - \underline{\mu})^T(\underline{x}_n - \underline{\mu})\Sigma^{-T}$$

$$\text{So } \frac{\partial}{\partial \Sigma}\ln p(X|\underline{\mu}, \Sigma) = -\frac{N}{2}\Sigma^{-T} + \frac{1}{2}\Sigma^{-T}\left(\sum_{n=1}^{N}(\underline{x}_n - \underline{\mu})(\underline{x}_n - \underline{\mu})^T\Sigma^{-T}\right) = 0$$

$$\Rightarrow N = \Sigma^{-T}\sum_{n=1}^{N}(\underline{x}_n - \underline{\mu})(\underline{x}_n - \underline{\mu})^T \Rightarrow \boxed{\Sigma_{ML} = \frac{1}{N}\sum_{n=1}^{N}(\underline{x}_n - \underline{\mu}_{ML})(\underline{x}_n - \underline{\mu}_{ML})^T}$$

But, when we check the expectation of our MLE. we have

$$\boxed{\mathbb{E}(\underline{\mu}_{ML}) = \frac{1}{N}\sum_{N}\mathbb{E}\,\underline{x}_n = \frac{1}{N}\cdot N\cdot\underline{\mu} = \underline{\mu}.}$$

$$\boxed{\begin{aligned}\mathbb{E}(\Sigma_{ML}) &= \frac{1}{N}\sum_{n}\mathbb{E}\left\{(\underline{x}_n - \frac{1}{N}\sum_{a}\underline{x}_a)(\underline{x}_n^T - \frac{1}{N}\sum_{b}\underline{x}_b^T)\right\} \\ &= \frac{1}{N}\sum_{n}\mathbb{E}\left\{\underline{x}_n\underline{x}_n^T - \frac{2}{N}\underline{x}_n\sum_{a}\underline{x}_a^T + \frac{1}{N^2}\sum_{a}\sum_{b}\underline{x}_a\underline{x}_b^T\right\} \\ &= \frac{1}{N}\left\{\underline{\mu}\underline{\mu}^T + \Sigma - 2\left(\underline{\mu}\underline{\mu}^T + \frac{1}{N}\Sigma\right) + \underline{\mu}\underline{\mu}^T + \frac{1}{N}\Sigma\right\} = \frac{N-1}{N}\Sigma\end{aligned}}$$

Biased Estimation.

We can construct an unbiased estimator

$$\hat{\Sigma} = \frac{1}{N-1} \sum_{n=1}^{N} (\underline{x}_n - \underline{\mu}_{ML})(\underline{x}_n - \underline{\mu}_{ML})^T$$

(5) Bayesian Inference.

Like Multinomial Distribution. we also want to find a conjugate prior for μ.

① Let's assume σ^2 is known. The likelihood function $p(\underline{X}|\mu)$ is

$$p(X|\mu) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_n-\mu)^2\right\}$$

If we treat $p(X|\mu)$ as a function of μ, it seems like it is a Gaussian.

So let's choose the prior to be another Gaussian

$$p(\mu) \sim N(\mu|\mu_0, \sigma_0^2).$$

Then

$$p(\mu|X) \propto p(X|\mu)\cdot p(\mu)$$

The exponent part for $p(X|\mu)\cdot p(\mu)$ is

$$-\frac{1}{2\sigma^2}\cdot\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{1}{2\sigma_0^2}(\mu-\mu_0)^2$$

$$= -\underbrace{\left(\frac{N}{2\sigma^2}+\frac{1}{2\sigma_0^2}\right)}_{\frac{1}{2\sigma_N^2}}\mu^2 + \underbrace{\left(\frac{1}{\sigma^2}\sum_{n=1}^{N}x_n + \frac{1}{\sigma_0^2}\mu_0\right)}_{\frac{\mu_N}{\sigma_N^2}}\mu + \text{const}$$

$$\mu_{ML} = \frac{1}{N}\sum_n x_n$$

↳ $\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0}$

↳ $\mu_N = \left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2}+\frac{1}{\sigma^2}\sum_{n=1}^{N}x_n\right)$

$$= \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{ML}$$

So, We find the posterior's μ_N & σ_N^2

② Let's assume the mean is known. we want to know the posterior for precision $\lambda = \frac{1}{\sigma^2}$

$$p(X|\lambda) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}} \cdot \sqrt{\lambda} \cdot \exp\{-\frac{\lambda}{2}(x_n-\mu)^2\}$$

$$\propto \lambda^{N/2} \cdot \exp\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n-\mu)^2\}$$

The corresponding prior is Gamma distribution (not 'gamma function')

similar

$$\text{Gamma}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

$$E(\lambda) = \frac{a}{b} \quad \text{Var}(\lambda) = \frac{a}{b^2}$$

So, the posterior will be (Given the prior to be $\text{Gamma}(\lambda|a_0,b_0)$)

$$p(\lambda|X) \propto \lambda^{a_0-1} \lambda^{N/2} \exp\{-b_0\lambda - \frac{\lambda}{2}\sum_n (x_n-\mu)^2\}$$

$\hookrightarrow \text{Gamma}(\lambda|a_N, b_N)$

$$a_N = a_0 + \frac{N}{2}$$

$$b_N = b_0 + \frac{1}{2}\sum(x_n-\mu)^2 = b_0 + \frac{N}{2}\sigma_{ML}^2$$