Probabilistic Model Setup and the MLE

Content:

1. Sample Mean and Sample Variance in Matrix Form
2. MLE of Multivariate Gaussian
3. Expectation and Variance of Gaussian Vectors (notes P8)
4. Probabilistic Model Setup and MLE of Parameters
5. Why MLE?

1 Sample Mean and Sample Variance in Matrix Form

Given a sample of random variables $\{X_n\} \overset{iid}{\sim}$dist, denoted by vector $\overrightarrow{X_n} \in \mathbb{R}^n$

$$Sample\ Mean\ (scalar): \qquad \overline{X_n} = \frac{1}{n}\overrightarrow{X_n}^T\overrightarrow{\mathbb{1}_n} \tag{1.1}$$

$$Mean\ Squared: \qquad \overline{X_n}^2 = \frac{1}{n^2}\overrightarrow{X_n}^T\mathbb{J}_{n\times n}\overrightarrow{X_n} \tag{1.2}$$

$$Sample\ Variance: \qquad \Sigma_{i=1}^n(x_i - \overline{X_n})^2 = \overrightarrow{X_n}^T(\mathbf{I}_n - \frac{1}{n}\mathbb{J})\overrightarrow{X_n} \tag{1.3}$$

Given a sample of random vectors $\{\overrightarrow{X_{p(n)}}\} \overset{iid}{\sim} dist$, denoted by matrix $X_{n\times p}$ with each row $\overrightarrow{X_{p(i)}}^T$.

$$Sample\ Mean\ (vector): \qquad \overline{\vec{X}_p} = \frac{1}{n}X_{n\times p}^T\overrightarrow{\mathbb{1}_n} \tag{1.4}$$

$$Sample\ Variance\ (matrix): \qquad \overline{\Sigma}_{p\times p} = X_{n\times p}^T(\mathbf{I}_n - \frac{1}{n}\mathbb{J})X_{n\times p} \tag{1.5}$$

Proof. 1.2
$\overline{X_n}^2 = \|\frac{1}{n}\overrightarrow{X_n}^T\overrightarrow{\mathbb{1}}_n\|^2 = \frac{1}{n^2}\overrightarrow{X_n}^T\overrightarrow{\mathbb{1}_n}\overrightarrow{\mathbb{1}_n}^T\overrightarrow{X_n} = \frac{1}{n^2}\overrightarrow{X_n}^T\mathbb{J}\overrightarrow{X_n}$ □

Proof. 1.3

$$\sum_{i=1}^n\left(x_i - \overline{X_n}\right)^2 = \|\overrightarrow{X_n} - \overline{X_n}\overrightarrow{\mathbb{1}_n}\|^2 = \overrightarrow{X_n}^T\overrightarrow{X_n} + n\overline{X_n}^2 - 2\overline{X_n}\overrightarrow{X_n}^T\overrightarrow{\mathbb{1}_n} = \overrightarrow{X_n}^T\overrightarrow{X_n} - n\overline{X_n}^2$$
$$= \overrightarrow{X_n}^T\overrightarrow{X_n} - n\frac{1}{n^2}\overrightarrow{X_n}^T\mathbb{J}\overrightarrow{X_n} = \overrightarrow{X_n}^T(\mathbf{I}_n - \frac{1}{n}\mathbb{J})\overrightarrow{X_n}$$

□

Proof. 1.5

$$\overline{\Sigma}_{p\times p} = \frac{1}{n}\sum_{i=1}^n(\overrightarrow{X_{p(i)}} - \overline{\vec{X}_p})(\overrightarrow{X_{p(i)}} - \overline{\vec{X}_p})^T = \frac{1}{n}(X_{n\times p}^T - \overline{\vec{X}_p}\overrightarrow{\mathbb{1}_n}^T)(X_{n\times p}^T - \overline{\vec{X}_p}\overrightarrow{\mathbb{1}_n}^T)^T$$
$$= \frac{1}{n}(X_{n\times p}^T - \frac{1}{n}X_{n\times p}^T\overrightarrow{\mathbb{1}_n}\overrightarrow{\mathbb{1}_n}^T)(X_{n\times p}^T - \frac{1}{n}X_{n\times p}^T\overrightarrow{\mathbb{1}_n}\overrightarrow{\mathbb{1}_n}^T)^T = \frac{1}{n}X_{n\times p}^T(\mathbf{I}_n - \frac{1}{n}\mathbb{J})(\mathbf{I}_n - \frac{1}{n}\mathbb{J})^T X_{n\times p}$$
$$= \frac{1}{n}X_{n\times p}^T(\mathbf{I}_n - \frac{2}{n}\mathbb{J} + \frac{1}{n^2}n\mathbb{J})X_{n\times p} = \boxed{\frac{1}{n}X_{n\times p}^T(\mathbf{I}_n - \frac{1}{n}\mathbb{J})X_{n\times p}}$$

□

2 MLE of Multivariate Gaussian

Fact 2.1. $\nabla_X ln|X| = \frac{1}{|X|}\nabla_X |X| = \frac{1}{|X|}|X|X^{-T} = X^{-T}$
Fact 2.2. $\nabla_X tr(AX) = \nabla_X tr(XA) = A^T$
Fact 2.3. *The quadratic form* $\vec{a}^T X\vec{a} = tr(X\vec{a}\vec{a}^T)$

Given $data = \{\overrightarrow{X_{p(n)}}\} \overset{iid}{\sim} N_P(\vec{\mu}_p, \Sigma_{p\times p})$

$$Likelihood:\ f_{X_{n\times p}}(X_{n\times p}) = \prod_{i=1}^{n}(2\pi)^{-\frac{p}{2}} \cdot |\Sigma|^{-\frac{1}{2}} \cdot exp\left\{-\frac{1}{2}(\overrightarrow{X_{p(i)}} - \overrightarrow{\mu_p})^T\Sigma^{-1}(\overrightarrow{X_{p(i)}} - \overrightarrow{\mu_p})\right\}$$

$$LogLikelihood:\ L_{X_{n\times p}}(X_{n\times p}) = -\frac{np}{2}ln2\pi + \frac{n}{2}ln|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{n}(\overrightarrow{X_{p(i)}} - \overrightarrow{\mu_p})^T\Sigma^{-1}_{p\times p}(\overrightarrow{X_{p(i)}} - \overrightarrow{\mu_p})$$

$$0 = \nabla_{\vec{\mu}_p} L_{X_{n\times p}}(X_{n\times p}) = -\frac{1}{2}\sum_{i=1}^{n}2\Sigma^{-1}_{p\times p}(\overrightarrow{X_{p(i)}} - \overrightarrow{\mu_p})(-\mathbf{I}_p) = \Sigma^{-1}_{p\times p}(\sum_{i=1}^{n}\overrightarrow{X_{p(i)}} - n\overrightarrow{\mu_p}) \implies$$

$$\widehat{\mu}_{p_{MLE}} = \frac{1}{n}\sum_{i=1}^{n}\overrightarrow{X_{p(i)}} = \boxed{\frac{1}{n}X^T_{n\times p}\overrightarrow{\mathbb{1}_n}}$$

$$0 = \nabla_{\Sigma^{-1}} L_{X_{n\times p}}(X_{n\times p}) = \nabla_{\Sigma^{-1}}\left\{\frac{n}{2}ln|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{n}tr\left[\Sigma^{-1}(\overrightarrow{X_{p(i)}} - \overrightarrow{\mu_p})(\overrightarrow{X_{p(i)}} - \overrightarrow{\mu_p})^T\right]\right\}$$

$$= \frac{n}{2}\Sigma_{p\times p} - \frac{1}{2}\sum_{i=1}^{n}(\overrightarrow{X_{p(i)}} - \overrightarrow{\mu_p})(\overrightarrow{X_{p(i)}} - \overrightarrow{\mu_p})^T \implies$$

$$\widehat{\Sigma}_{MLE} = \frac{1}{n}\sum_{i=1}^{n}(\overrightarrow{X_{p(i)}} - \overrightarrow{\mu_p})(\overrightarrow{X_{p(i)}} - \overrightarrow{\mu_p})^T = \boxed{\frac{1}{n}X^T_{n\times p}(\mathbf{I}_n - \frac{1}{n}\mathbb{J})X_{n\times p}}$$

3 to be added

4 Probabilistic Model Setup and the MLE of Parameters

4.1 Model Setup

$$\overrightarrow{Y_n} = X_{n\times p}\overrightarrow{\beta_p} + \overrightarrow{\epsilon_n}$$

Assumptions:

1. $\vec{\epsilon}_n \sim N_n(\vec{0}_n, \sigma^2\mathbf{I}_n)$, where σ^2 is known.
2. For $i = 1, ..., n$, $\overrightarrow{X_{p(i)}}^T \overset{iid}{\sim}$ some known dist with pdf $f_{\vec{X}_p}(\overrightarrow{\xi_p})$
3. $\vec{\epsilon}_n \perp\!\!\!\perp X_{n\times p}$, or equivalently $\overrightarrow{X_{p_{(i)}}} \perp\!\!\!\perp \overrightarrow{\epsilon_j} \quad \forall\, i, j = 1, ..., n$ (easily violated)

MLE: Given $data = (\overrightarrow{y_n}, X_{n\times p}),\ \overrightarrow{Y_n}|X_{n\times p} \sim N_n(X_{n\times p}\overrightarrow{\beta_p}, \sigma^2\mathbf{I}_n)$

$Likelihood:\ f_{\vec{\theta}}(\overrightarrow{Y_n}, X_{n\times p}|\vec{\theta}) = f_{\overrightarrow{Y_n}}(\overrightarrow{Y_n}|X_{n\times p}, \vec{\theta}) \cdot f_{X_{n\times p}}(X_{n\times p}|\vec{\theta}) = (2\pi)^{-\frac{n}{2}} \cdot (\sigma^2)^{-\frac{n}{2}} \cdot$

$exp\left\{-\frac{1}{2\sigma^2}(\overrightarrow{Y_n} - X\overrightarrow{\beta_p})^T(\overrightarrow{Y_n} - X\overrightarrow{\beta_p})\right\}\prod_{i=1}^{n} f_{\vec{X}_{p_{(i)}}}(\vec{X}_{p_{(i)}})$

$LogLikelihood: L_{\vec{\theta}}(\overrightarrow{Y_n}, X_{n\times p}|\vec{\theta}) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}(\overrightarrow{Y_n} - X\overrightarrow{\beta_p})^T(\overrightarrow{Y_n} - X\overrightarrow{\beta_p}) +\ const.$

set $0 = \nabla_{\vec{\beta}} L_{\vec{\theta}}(\overrightarrow{Y_n}, X_{n\times p}|\vec{\theta}) = -\frac{1}{2\sigma^2} \cdot 2(-X^T_{n\times p})(\overrightarrow{Y_n} - X\overrightarrow{\beta_p}) \Rightarrow X^T_{n\times p}\overrightarrow{Y_n} = X^TX\overrightarrow{\beta_p} \Longrightarrow$

$\boxed{\widehat{\overrightarrow{\beta}}_{p\,MLE} = (X^TX)^{-1}X^T\overrightarrow{Y_n}}$

set $0 = \nabla_{\sigma^2} L_{\vec{\theta}}(\overrightarrow{Y_n}, X_{n\times p}|\vec{\theta}) = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\|\overrightarrow{Y_n} - X_{n\times p}\overrightarrow{\beta_p}\|^2 \Rightarrow \boxed{\widehat{\sigma^2}_{MLE} = \frac{1}{n}\|\overrightarrow{Y_n} - X\overrightarrow{\beta_p}\|^2 = \frac{1}{n}\left\|(\mathbf{I}-\mathbf{P})\overrightarrow{Y_n}\right\|^2}$

$= \frac{1}{n}\overrightarrow{Y_n}^T(\mathbf{I}-\mathbf{P})\overrightarrow{Y_n} \sim \frac{1}{n}\sigma^2\chi^2_{n-p}$

Note:

1. Even if in assumption 2 the rows of design matrix X are dependent, $\hat{\vec{\beta}}_{MLE} = \hat{\vec{\beta}}_{LSE} =$ sample mean is still true.
2. Even if in assumption 1, the distribution is not Gaussian, it is asymptotic Gaussian. $\hat{\vec{\beta}}_{MLE} \xrightarrow{D} N_p(\beta_p, \frac{I_\beta^{-1}}{n})$
3. If assumption 3 is violated, the model setup is no longer valid.

4.2 Non-i.i.d errors

If assumption 1 is changed to $\overrightarrow{\epsilon_n} \sim N_n(\vec{0}_n, \text{some known } \Sigma)$ (non i.i.d. errors), then
MLE: given $data = (\overrightarrow{y_n}, X_{n\times p}),\ \overrightarrow{Y_n}|X_{n\times p} \sim N_n(X_{n\times p}\overrightarrow{\beta_p}, \Sigma_{p\times p})$

$Likelihood:\ f_{\vec{\theta}}(\overrightarrow{Y_n}, X_{n\times p}|\vec{\theta}) = f_{\overrightarrow{Y_n}}(\overrightarrow{Y_n}|X_{n\times p}, \vec{\theta}) \cdot f_{X_{n\times p}}(X_{n\times p}|\vec{\theta}) = (2\pi)^{-\frac{n}{2}} \cdot |\Sigma|^{-\frac{1}{2}} \cdot$

$exp\left\{-\frac{1}{2}(\overrightarrow{Y_n} - X\overrightarrow{\beta_p})^T\Sigma^{-1}_{p\times p}(\overrightarrow{Y_n} - X\overrightarrow{\beta_p})\right\}\prod_{i=1}^{n} f_{\vec{X}_{p_{(i)}}}(\vec{X}_{p_{(i)}})$

$LogLikelihood: L_{\vec{\theta}}(\overrightarrow{Y_n}, X_{n\times p}|\vec{\theta}) = -\frac{n}{2}ln(2\pi) + \frac{1}{2}ln|\Sigma^{-1}| - \frac{1}{2}(\overrightarrow{Y_n} - X\overrightarrow{\beta_p})^T\Sigma^{-1}(\overrightarrow{Y_n} - X\overrightarrow{\beta_p}) +\ const.$

set $0 = \nabla_{\vec{\beta}} L_{\vec{\theta}}(\overrightarrow{Y_n}, X_{n\times p}|\vec{\theta}) = -\frac{1}{2} \cdot 2(-X^T_{n\times p})\Sigma^{-1}(\overrightarrow{Y_n} - X\overrightarrow{\beta_p}) \Rightarrow X^T_{n\times p}\Sigma^{-1}\overrightarrow{Y_n} = X^T\Sigma^{-1}X\overrightarrow{\beta_p} \Longrightarrow$

$\boxed{\widehat{\overrightarrow{\beta}}_{p\,MLE} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}\overrightarrow{Y_n}}$

set $0 = \nabla_{\Sigma^{-1}} L_{\vec{\theta}}(\overrightarrow{Y_n}, X_{n\times p}|\vec{\theta}) = \frac{1}{2}\Sigma^T - \frac{1}{2}\nabla_{\Sigma^{-1}} tr\left[\Sigma^{-1}(\overrightarrow{Y_n} - X_{n\times p}\overrightarrow{\beta_p}) \cdot (\overrightarrow{Y_n} - X_{n\times p}\overrightarrow{\beta_p})^T\right]$

$= \frac{1}{2}\Sigma^T - \frac{1}{2}(\overrightarrow{Y_n} - X_{n\times p}\overrightarrow{\beta_p}) \cdot (\overrightarrow{Y_n} - X_{n\times p}\overrightarrow{\beta_p})^T \Rightarrow \boxed{\hat{\Sigma}_{MLE} = (\overrightarrow{Y_n} - X\widehat{\overrightarrow{\beta}}_{p\,MLE}) \cdot (\overrightarrow{Y_n} - X\widehat{\overrightarrow{\beta}}_{p\,MLE})^T}$

5 Why MLE

No matter what distribution $\vec{\epsilon}_n$ has, under the assumption that data are independent and identically distributed, MLE is asymptotically normal and efficient (unbiased and hits the Cramer-Rao bound, hence has the MV) when $n \to \infty$, thus the best estimator. $\hat{\theta}_{MLE} \xrightarrow{D} N(\theta, \frac{\mathbf{I}^{-1}}{n})$ where $\mathbf{I}$ is Fisher Information Matrix (see summary notes of Chapter 8)
Note: How large is large? Empirically, $n \geq 50p$ for the design matrix.

Proof. We need Theorem 1,2,3 in Notes of Chapter 6 Part II, Limit Theorems in Multivariate Case. Under the same assumptions in 4.1 except that assumption 1 is modified to $\epsilon_i \overset{iid}{\sim}$ some unknown dist. with known mean 0 and variance σ^2, then

$$\widehat{\overrightarrow{\beta}}_{MLE} - \overrightarrow{\beta} = (X^TX)^{-1}X^T\overrightarrow{y} - \overrightarrow{\beta} = (X^TX)^{-1}X^T\overrightarrow{\epsilon_n}$$

$$\frac{1}{n}X^T_{n\times p}X_{n\times p} \xrightarrow[(LLN)]{P} E(\overrightarrow{X_p}\overrightarrow{X_p}^T) = \widetilde{\Sigma}_{p\times p} \qquad ①$$

$$X^T_{n\times p}\overrightarrow{\epsilon_n} = \sum_{i=1}^{n}\overrightarrow{X_{p(i)}}\cdot\epsilon_i \xrightarrow[(CLT)]{D} N_p\left(\sum_{i=1}^{n}E(\overrightarrow{X_{p(i)}}\cdot\epsilon_i), \sum_{i=1}^{n}Var(\overrightarrow{X_{p(i)}}\cdot\epsilon_i)\right) \qquad ②$$

but $E\left(\overrightarrow{X_{p(i)}}\cdot\epsilon_i\right) = E\left(\overrightarrow{X_{p(i)}}\right)\cdot E(\epsilon_i) = \vec{0}_p$

$$Var\left(\overrightarrow{X_{p(i)}}\cdot\epsilon_i\right) = E\left[\left(\overrightarrow{X_{p(i)}}\cdot\epsilon_i\right)\left(\overrightarrow{X_{p(i)}}\cdot\epsilon_i\right)^T\right] - E\left(\overrightarrow{X_{p(i)}}\cdot\epsilon_i\right)E\left(\overrightarrow{X_{p(i)}}\cdot\epsilon_i\right)^T = E\left(\epsilon_i^2\overrightarrow{X_{p(i)}}\overrightarrow{X_{p(i)}}^T\right)$$

$$= \sigma^2\widetilde{\Sigma}_{p\times p} \overset{②}{\Longrightarrow} X^T_{n\times p}\overrightarrow{\epsilon_n} \xrightarrow[CLT]{D} N_p(\vec{0}_p, n\sigma^2\widetilde{\Sigma}_{p\times p}) \qquad ③$$

$$①^{-1} \times ③ \Rightarrow n(X^TX)^{-1}X^T\overrightarrow{\epsilon_n} \xrightarrow[\text{(Slutsky's Theorem)}]{D} N_p(\vec{0}_p, n\sigma^2\widetilde{\Sigma}^{-1}_{p\times p}) \Rightarrow$$

$$\boxed{(X^TX)^{-1}X^T\overrightarrow{\epsilon_n} \xrightarrow{D} N_p(\vec{0}_p, \frac{\sigma^2}{n}\widetilde{\Sigma}^{-1}_{p\times p})}\text{, where } \widetilde{\Sigma}_{p\times p} = E(\overrightarrow{X_p}\overrightarrow{X_p}^T)$$

□