Maximum Likelihood Estimate of Multivariate Normal Distribution

Instructor : Dr. Arabin Kumar Dey

1 Some Results on Vector and Matrix Calculus

1.1 Definition of derivative with respect to Vector

Suppose, $f : R^n \to R$, e.g., $f(\underline{x}) = \underline{x}^T A \underline{x}$ where $\underline{x}$ is a column vector of dimension $n \times 1$. Derivative of f with respect to $\underline{x}$ i.e.

$$\frac{df(\underline{x})}{d\underline{x}} = [\frac{\delta f(\underline{x})}{\delta x_1}, \frac{\delta f(\underline{x})}{\delta x_2}, \cdots, \frac{\delta f(\underline{x})}{\delta x_n}]^T$$

is a row vector of dimension $1 \times n$. $h\underline{e}_i = \underline{h}$.

$$\frac{\delta f(\underline{x})}{\delta x_i} = \lim_{||\underline{h}|| \to \infty} \frac{f(\underline{x} + h\underline{e}_i) - f(\underline{x})}{||\underline{h}||}$$

1.2 Definition of derivative with respect to Matrix

Suppose $f : R^{m \times n} \to R$. Derivative of f with respect to A (a $m \times n$ matrix) is of dimension $n \times m$ can be written as $\frac{df(A)}{dA} = ((\frac{\delta f(A)}{\delta a_{ji}}))$.

1.2 Quadratic forms

1.2.1 Basic quadratic form

Now consider something slightly more complicated. Let A be a symmetric $n \times n$ matrix

$$f(x) = x^T A x$$

Now we have a function from $\mathbb{R}^n \to \mathbb{R}$ and the derivative should be an $1 \times n$ matrix (it also maps $\mathbb{R}^n \to \mathbb{R}$).

We can also use first principles in calculating the partial derivatives.

Do some calculations by taking

$$\frac{(x + he_i)^T A(x + he_i) - x^T A x}{h} = x^T A^T he_i + x^T A he_i + h^2 e_i^T A e_i$$

Since A is symmetric, we get

$$\frac{dx^T A x}{dx} = x^T A^T + x^T A = x^T A + x^T A = 2x^T A$$

(We could also have just used the product rule to derive this)

Figure 1: Vector Calculus

1.3 Some Specific Examples

1.4 Some Specific Useful Results

1.5 Result 1 :

Suppose, A is matrix of dimension $n \times m$ and B is a matrix of dimension $m \times n$. Then, $\frac{dtr(AB)}{dA} = B$.

Suppose, $\frac{dtr(AB)}{dA} = \frac{d\sum_{k=1}^{n}\sum_{j'=1}^{m} a_{kj'} b_{j'k}}{dA} = ((\frac{d\sum_{k=1}^{n}\sum_{j'=1}^{m} a_{kj'} b_{j'k}}{da_{ji}})) = ((b_{ij})) = B$

1.6 Result 2 :

Suppose A is symmetric matrix, $\frac{d\ln|A|}{dA} = A^{-1}$. Since A is symmetric, $A = A^T$ implies $|A| = |A^T|$. We know that $|A^T| = \sum_{i=1}^{n}(-1)^{i+j}a_{ji}M_{ji}$ where, M_{ji} is the minor corresponding to the j-th row and i-th column. Taking derivative with respect to a_{ji}, $\frac{\delta|A^T|}{\delta a_{ji}} = (-1)^{i+j}M_{ji}$. Let's assume $\tilde{A}$ is the adjugate matrix.

$$\frac{d\ln|A|}{dA} = ((\frac{\delta\ln|A|}{\delta a_{ji}})) = ((\frac{\delta\ln|A^T|}{\delta a_{ji}})) = \frac{1}{|A^T|}(((-1)^{i+j}M_{ji})) = .\frac{\tilde{A}}{|A|} = A^{-1}$$

2 Derivation of MLE for Multivariate Normal Distribution

4 MLE

We now put this all together

$$l(\mu, \Sigma|\mathcal{D}) = -\frac{Nd}{2} log(2\pi) - \frac{N}{2} log|\Sigma| + \sum_{i=1}^{N} (x_i - \mu)^T \Sigma^{-1} (x_i - \mu)/2$$

Take the derivative wrt μ

$$\frac{\partial l}{\partial \mu} = \sum_{i=1}^{N} (x_i - \mu)^T \Sigma^{-1}$$

(Take $B = \Sigma^{-1}$, $A = I$ from our prev calculation.)

setting to 0, multiplying both sides by Σ and we get our MLE for μ

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

Figure 2: MLE for mean

4.1 MLE for Σ

We first use the fact that $|A^{-1}| = |A|^{-1}$ and the trace trick.

$$l = \frac{N}{2} log|\Sigma^{-1}| - \sum_{i=1}^{N} tr\left(\Sigma^{-1}(x_i - \mu)(x_i - \mu)^T\right)/2 + C$$

Taking derivatives wrt to Σ^{-1} we get

$$\frac{\partial l}{\partial \Sigma^{-1}} = \frac{N}{2}\Sigma - \sum_{i=1}^{N}(x_i - \mu)(x_i - \mu)^T/2$$

(Setting to 0, using our MLE $\hat{\mu}$ and doing some algebra gives)

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \hat{\mu})(x_i - \hat{\mu})^T$$

Figure 3: MLE for Variance Covariance Matrix