The Multivariate Normal Distribution

Sampling from a Multivariate Normal Distribution
MLE, Sampling Distribution of $\bar{X}$ and S

Santiago Alférez

Agosto de 2020

Análisis Estadístico de Datos
MACC

Universidad del Rosario

Contenidos

The Multivariate Normal Likelihood

Maximum Likelihood Estimation of μ and Σ

The Sampling Distribution of
$\overline{\mathbf{X}}$ and $\mathbf{S}$

Large-Sample Behavior of $\overline{X}$ and S

The Multivariate Normal Likelihood

The Multivariate Normal Likelihood

The multivariate normal likelihood

- Let us assume that the $p \times 1$ vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ represent a random sample from a multivariate normal population with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.
- Since $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ are mutually independent and each has distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the joint density function of all the observations is the product of the marginal normal densities:

$$\left\{ \begin{array}{c} \text{Joint density} \\ \text{of } \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \end{array} \right\} = \prod_{j=1}^{n} \left\{ \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} e^{-(\mathbf{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu})/2} \right\}$$

$$= \frac{1}{(2\pi)^{np/2}} \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} e^{-\sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu})/2}$$

The Multivariate Normal Likelihood

The multivariate normal likelihood

- When the numerical values of the observations become available, they may be substituted for the $\mathbf{x}_j$ in the above Equation.
- The resulting expression, now considered as a function of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for the fixed set of observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, is called the likelihood.
- Many good statistical procedures employ values for the population parameters that "best" explain the observed data.
- One meaning of best is to select the parameter values that maximize the joint density evaluated at the observations. This technique is called maximum likelihood estimation, and the maximizing parameter values are called maximum likelihood estimates.

The Multivariate Normal Likelihood

- We shall consider maximum likelihood estimation of the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ for a multivariate normal population.
- To do so, we take the observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ as fixed and consider the joint density of Equation evaluated at these values. The result is the likelihood function.

The trace of a symmetric matrix

Let $\mathbf{A}$ be a $k \times k$ symmetric matrix and $\mathbf{x}$ be a $k \times 1$ vector. Then

a $\mathbf{x}'\mathbf{A}\mathbf{x} = \operatorname{tr}(\mathbf{x}'\mathbf{A}\mathbf{x}) = \operatorname{tr}(\mathbf{A}\mathbf{x}\mathbf{x}')$

b $\operatorname{tr}(\mathbf{A}) = \sum_i^k \lambda_i$, where the λ_i are the eigenvalues of $\mathbf{A}$

The Multivariate Normal Likelihood

$$\left\{ \begin{array}{c} \text{Joint density} \\ \text{of } \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \end{array} \right\} = \frac{1}{(2\pi)^{np/2}} \frac{1}{|\boldsymbol{\Sigma}|^{n/2}} e^{-\sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu})/2}$$

Applying the trace of a symmetric matrix

$$(\mathbf{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu}) = \operatorname{tr}\left[(\mathbf{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu})\right] = \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu}) (\mathbf{x}_j - \boldsymbol{\mu})'\right]$$

Since the trace of a sum of matrices is equal to the sum of the traces of the matrices:

$$\begin{aligned} \sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu}) &= \sum_{j=1}^{n} \operatorname{tr}\left[(\mathbf{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu})\right] \\ &= \sum_{j=1}^{n} \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu}) (\mathbf{x}_j - \boldsymbol{\mu})'\right] \\ &= \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1} \left(\sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu}) (\mathbf{x}_j - \boldsymbol{\mu})' \right)\right] \end{aligned}$$

The Multivariate Normal Likelihood

$$\sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu}) = \text{tr}\left[\boldsymbol{\Sigma}^{-1} \left(\sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu})(\mathbf{x}_j - \boldsymbol{\mu})'\right)\right]$$

Adding and subtracting $\overline{\mathbf{x}} = (1/n)\sum_{j=1}^{n} \mathbf{x}_j$ in each $(\mathbf{x}_j - \boldsymbol{\mu})$

Because the cross-product terms, $\sum_{j=1}^{n} (\mathbf{x}_j - \overline{\mathbf{x}})(\overline{\mathbf{x}} - \boldsymbol{\mu})'$ and $\sum_{j=1}^{n} (\overline{\mathbf{x}} - \boldsymbol{\mu})(\mathbf{x}_j - \overline{\mathbf{x}})'$ are both matrices of zeros:

$$\begin{aligned}
&\sum_{j=1}^{n} (\mathbf{x}_j - \overline{\mathbf{x}} + \overline{\mathbf{x}} - \boldsymbol{\mu})(\mathbf{x}_j - \overline{\mathbf{x}} + \overline{\mathbf{x}} - \boldsymbol{\mu})' \\
&= \sum_{j=1}^{n} (\mathbf{x}_j - \overline{\mathbf{x}})(\mathbf{x}_j - \overline{\mathbf{x}})' + \sum_{j=1}^{n} (\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})' \\
&= \sum_{j=1}^{n} (\mathbf{x}_j - \overline{\mathbf{x}})(\mathbf{x}_j - \overline{\mathbf{x}})' + n(\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})'
\end{aligned}$$

The Multivariate Normal Likelihood

$$\sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x}_j - \boldsymbol{\mu}) = \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1} \left(\sum_{j=1}^{n} (\mathbf{x}_j - \boldsymbol{\mu}) (\mathbf{x}_j - \boldsymbol{\mu})'\right)\right]$$
$$= \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1} \left(\sum_{j=1}^{n} (\mathbf{x}_j - \overline{\mathbf{x}}) (\mathbf{x}_j - \overline{\mathbf{x}})' + n(\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})'\right)\right]$$

Join density of a random sample from a multivariate normal population

$$\left\{ \begin{array}{c} \text{Joint density of} \\ \mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n \end{array} \right\} = (2\pi)^{-np/2} |\boldsymbol{\Sigma}|^{-n/2}$$
$$\times \exp\left\{- \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1} \left(\sum_{j=1}^{n} (\mathbf{x}_j - \overline{\mathbf{x}}) (\mathbf{x}_j - \overline{\mathbf{x}})' + n(\overline{\mathbf{x}} - \boldsymbol{\mu})(\overline{\mathbf{x}} - \boldsymbol{\mu})'\right)\right] /2\right\}$$

The Multivariate Normal Likelihood

Likelihood function from the join density

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} e^{-\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\overline{\mathbf{x}})(\mathbf{x}_j-\overline{\mathbf{x}})'+n(\overline{\mathbf{x}}-\boldsymbol{\mu})(\overline{\mathbf{x}}-\boldsymbol{\mu})'\right)\right]/2}$$

Different ways of the exponent in the likelihood function

$$\begin{aligned}
&\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}\left(\mathbf{x}_j-\overline{\mathbf{x}}\right)\left(\mathbf{x}_j-\overline{\mathbf{x}}\right)'+n(\overline{\mathbf{x}}-\boldsymbol{\mu})(\overline{\mathbf{x}}-\boldsymbol{\mu})'\right)\right] \\
&= \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}\left(\mathbf{x}_j-\overline{\mathbf{x}}\right)\left(\mathbf{x}_j-\overline{\mathbf{x}}\right)'\right)\right]+n\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})(\overline{\mathbf{x}}-\boldsymbol{\mu})'\right] \\
&= \operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^{n}\left(\mathbf{x}_j-\overline{\mathbf{x}}\right)\left(\mathbf{x}_j-\overline{\mathbf{x}}\right)'\right)\right]+n(\overline{\mathbf{x}}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})
\end{aligned}$$

Maximum Likelihood Estimation of μ and Σ

Maximum Likelihood Estimation of μ and Σ

Inequality to obtain the MLE of μ and Σ

Given a $p \times p$ symmetric positive definite matrix $\mathbf{B}$ and a scalar $b > 0$, it follows that

$$\frac{1}{|\mathbf{\Sigma}|^b} e^{-\operatorname{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{B}\right)/2} \leq \frac{1}{|\mathbf{B}|^b}(2b)^{pb} e^{-bp}$$

for all positive definite $\underset{(p\times p)}{\mathbf{\Sigma}}$, with equality holding only for $\mathbf{\Sigma} = (1/2b)\mathbf{B}$

Maximum Likelihood Estimation of μ and Σ

MLE of μ and Σ

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from a normal population with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$. Then

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{X}} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \frac{1}{n} \sum_{j=1}^{n} \left(\mathbf{X}_j - \overline{\mathbf{X}}\right) \left(\mathbf{X}_j - \overline{\mathbf{X}}\right)' = \frac{(n-1)}{n} \mathbf{S}$$

are the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively. Their observed values, $\overline{\mathbf{x}}$ and $(1/n) \sum_{j=1}^{n} \left(\mathbf{x}_j - \overline{\mathbf{x}}\right) \left(\mathbf{x}_j - \overline{\mathbf{x}}\right)'$, are called the maximum likelihood estimates of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.

Maximum Likelihood Estimation of μ and Σ

The exponent in the likelihood function without the $-\frac{1}{2}$

$$\operatorname{tr}\left[\mathbf{\Sigma}^{-1}\left(\sum_{j=1}^{n}\left(\mathbf{x}_j-\overline{\mathbf{x}}\right)\left(\mathbf{x}_j-\overline{\mathbf{x}}\right)'\right)\right]+n(\overline{\mathbf{x}}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})$$

Proof of MLE of μ and Σ

Since $\mathbf{\Sigma}^{-1}$ is positive definite, so the distance $(\overline{\mathbf{x}}-\boldsymbol{\mu})'\mathbf{\Sigma}^{-1}(\overline{\mathbf{x}}-\boldsymbol{\mu})>0$ unless $\boldsymbol{\mu}=\overline{\mathbf{x}}$. Thus, the likelihood is maximized with respect to $\boldsymbol{\mu}$ at $\hat{\boldsymbol{\mu}}=\overline{\mathbf{x}}$.

Then, we need only maximize

$$L(\boldsymbol{\mu},\mathbf{\Sigma})=\frac{1}{(2\pi)^{np/2}|\mathbf{\Sigma}|^{n/2}}e^{-\operatorname{tr}\left[\mathbf{\Sigma}^{-1}\left(\sum_{j=1}^{n}(\mathbf{x}_j-\overline{\mathbf{x}})(\mathbf{x}_j-\overline{\mathbf{x}})'\right)\right]/2}$$

with respect to Σ.

Maximum Likelihood Estimation of μ and Σ

Given a $p \times p$ symmetric positive definite matrix $\mathbf{B}$ and a scalar $b > 0$, it follows that $\frac{1}{|\boldsymbol{\Sigma}|^b} e^{-\operatorname{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{B}\right)/2} \leq \frac{1}{|\mathbf{B}|^b}(2b)^{pb}e^{-bp}$ for all positive definite $\underset{(p\times p)}{\boldsymbol{\Sigma}}$, with equality holding only for $\boldsymbol{\Sigma} = (1/2b)\mathbf{B}$

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} e^{-\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}\left(\sum_{j=1}^n (\mathbf{x}_j - \overline{\mathbf{x}})(\mathbf{x}_j - \overline{\mathbf{x}})'\right)\right]/2}$$

Proof of MLE of μ and Σ

Using the result of the blue block with $b = n/2$ and $\mathbf{B} = \sum_{j=1}^n (\mathbf{x}_j - \overline{\mathbf{x}})(\mathbf{x}_j - \overline{\mathbf{x}})'$, the maximum occurs at $\hat{\boldsymbol{\Sigma}} = (1/n)\sum_{j=1}^n (\mathbf{x}_j - \overline{\mathbf{x}})(\mathbf{x}_j - \overline{\mathbf{x}})'$

- The maximum likelihood estimators are random quantities. They are obtained by replacing the observations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ in the expressions for $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ with the corresponding random vectors, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$

Maximum Likelihood Estimation of μ and Σ

Relation betweek likelihood and generalized variance

- The maximum likelihood estimator $\overline{\mathbf{X}}$ is a random vector and the maximum likelihood estimator $\hat{\mathbf{\Sigma}}$ is a random matrix. The maximum likelihood estimates are their particular values for the given data set.
- The maximum of the likelihood is $L(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}) = \frac{1}{(2\pi)^{np/2}} e^{-np/2} \frac{1}{|\hat{\mathbf{\Sigma}}|^{n/2}}$ or, since $|\hat{\mathbf{\Sigma}}| = [(n-1)/n]^p |\mathbf{S}|$

$$L(\hat{\boldsymbol{\mu}}, \hat{\mathbf{\Sigma}}) = \text{ constant } \times (\text{ generalized variance })^{-n/2}$$

- The generalized variance determines the **peakedness** of the likelihood function and, consequently, is a natural measure of variability when the parent population is multivariate normal.

Maximum Likelihood Estimation of μ and Σ

Invariance property of MLE

Let $\hat{\boldsymbol{\theta}}$ be the MLE of $\boldsymbol{\theta}$, and consider estimating the parameter $h(\boldsymbol{\theta})$ which is a function of $\boldsymbol{\theta}$. Then the MLE of

$$\underset{\text{(a function of } \theta)}{h(\theta)} \quad \text{is given by} \quad \underset{\text{(a function of } \hat{\theta})}{h(\hat{\theta})}$$

1. The maximum likelihood estimator of $\boldsymbol{\mu}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$ is $\hat{\boldsymbol{\mu}}'\hat{\boldsymbol{\Sigma}}^{-1}\hat{\boldsymbol{\mu}}$, where $\hat{\boldsymbol{\mu}} = \overline{\mathbf{X}}$ and $\hat{\boldsymbol{\Sigma}} = ((n-1)/n)\mathbf{S}$ are the maximum likelihood estimators of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ respectively.
2. The maximum likelihood estimator of $\sqrt{\sigma_{ii}}$ is $\sqrt{\hat{\sigma}_{ii}}$, where
$$\hat{\sigma}_{ii} = \frac{1}{n}\sum_{j=1}^{n}\left(X_{ij} - \bar{X}_i\right)^2$$
is the maximum likelihood estimator of $\sigma_{ii} = \operatorname{Var}\left(X_i\right)$

Maximum Likelihood Estimation of μ and Σ

Sufficient Statistics

- The sample estimates $\overline{\mathbf{X}}$ and $\mathbf{S}$ are sufficient statistics
- This means that all of the information contained in the data can be summarized by these two statistics alone
- This is only true if the data follow a multivariate normal distribution - if they do not, other terms are needed (i.e., skewness array, kurtosis array, etc...)
- Some statistical methods only use one or both of these matrices in their analysis procedures and not the actual data

The Sampling Distribution of $\overline{\mathrm{X}}$ and S

Some considerations

- With $(p=1)$, we know that X is normal with mean $\mu=$ (population mean) and variance $\frac{1}{n} \sigma^2=\frac{\text{population variance}}{\text{sample size}}$
- The result for the multivariate case $(p \geq 2)$ is analogous in that $\overline{\mathbf{X}}$ has a normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $(1/n) \mathbf{\Sigma}$.
- For the sample variance, recall that $(n-1) s^2=\sum_{j=1}^n\left(X_j-\bar{X}\right)^2$ is distributed as σ^2 times a chi-square variable having $n-1$ df.
- This chi-square is the distribution of a sum of squares of independent standard normal random variables. That is, $(n-1) s^2$ is distributed as $\sigma^2\left(Z_1^2+\cdots+Z_{n-1}^2\right)=\left(\sigma Z_1\right)^2+\cdots+\left(\sigma Z_{n-1}\right)^2$.
- The individual terms σZ_i are independently distributed as $N\left(0, \sigma^2\right)$. It is this latter form that is suitably generalized to the basic sampling distribution for the sample covariance matrix.

The Sampling Distribution of $\overline{\mathrm{X}}$ and S

MLE of μ and Σ

The maximum likelihood estimator of μ is $\hat{\mu} = \overline{\mathbf{X}} = \frac{1}{n}\sum_{j=1}^{n} X_j$ and the ML estimator of $\boldsymbol{\Sigma}$ is

$$\hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\mathbf{S} = \mathbf{S}_n = \frac{1}{n}\sum_{j=1}^{n}\left(\mathbf{X}_j - \hat{\boldsymbol{\mu}}\right)\left(\mathbf{X}_j - \hat{\boldsymbol{\mu}}\right)'$$

Sampling distribution of $\hat{\mu}$

The estimator is a linear combination of normal random vectors each from $\mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ i.i.d.

$$\hat{\boldsymbol{\mu}} = \overline{\mathbf{X}} = \frac{1}{n}\mathbf{X}_1 + \frac{1}{n}\mathbf{X}_2 + \cdots + \frac{1}{n}\mathbf{X}_n$$

So $\hat{\mu} = \overline{X}$ also has a normal distribution $N_p(\boldsymbol{\mu}, (1/n)\boldsymbol{\Sigma})$

The Sampling Distribution of $\overline{\mathrm{X}}$ and S

$$\hat{\boldsymbol{\Sigma}} = \frac{n-1}{n}\mathbf{S}$$

Sampling distribution of $\hat{\Sigma}$

The matrix

$$(n-1)\mathbf{S} = \sum_{j=1}^{n} (\mathbf{x}_j - \overline{\mathbf{x}})(\mathbf{x}_j - \overline{\mathbf{x}})'$$

is distributed as a Wishart random matrix with $(n-1)$ degrees of freedom.

Whishart distribution

- A multivariate analogue to the chi-square distribution.
- It's defined as

$$\begin{aligned} W_m(\cdot \mid \boldsymbol{\Sigma}) &= \text{Wishart distribution with } m \text{ degrees of freedom} \\ &= \text{The distribution of } \sum_{j=1}^{m} \mathbf{Z}_j \mathbf{Z}_j' \end{aligned}$$

where $\mathbf{Z}_j \sim \mathcal{N}_p(\mathbf{0}, \boldsymbol{\Sigma})$ and independent.

Note: $\overline{\mathrm{X}}$ and S are independent.

Large-Sample Behavior of $\overline{X}$ and S

Law of Large Numbers

Data are not always (multivariate) normal

The law of large numbers for multivariate data

- Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent observations from a population with mean $E(\mathbf{X}) = \boldsymbol{\mu}$.
- Then $\overline{\mathbf{X}} = (1/n)\sum_{j=1}^{n} \mathbf{X}_j$ converges in probability to μ as n gets large; that is,

$$\overline{\mathbf{X}} \rightarrow \mu \text{ for large samples}$$

- And

$$\mathbf{S}\,(\text{or } \mathbf{S}_n) \text{ approach } \boldsymbol{\Sigma} \text{ for large samples}$$

- These are true regardless of the true distribution of the $\mathbf{X}_j$'s.

Central Limit Theorem

The central limit theorem

- Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be independent observations from a population with mean $E(\mathbf{X}) = \boldsymbol{\mu}$ and finite (non-singular, full rank), covariance matrix $\boldsymbol{\Sigma}$.
- Then $\sqrt{n}(\mathbf{X} - \boldsymbol{\mu})$ has an approximate $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ distribution if $n >> p$ (i.e., "much larger than").
- So, for "large" n

$$\overline{\mathbf{X}} = \text{ Sample mean vector } \approx \mathcal{N}\left(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma}\right)$$

regardless of the underlying distribution of the $\mathbf{X}_j$'s.

What if Σ is unknown?

If n is large "enough", S will be close to Σ, So

$$\sqrt{n}(\overline{\mathbf{X}} - \boldsymbol{\mu}) \approx \mathcal{N}_p(\mathbf{0}, \mathbf{S}) \text{ or } \overline{\mathbf{X}} \approx \mathcal{N}_p\left(\boldsymbol{\mu}, \frac{1}{n}\mathbf{S}\right)$$

since $n(\overline{\mathbf{X}} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\overline{\mathbf{X}} - \boldsymbol{\mu}) \sim \chi^2_p$

$$n(\overline{\mathbf{X}} - \boldsymbol{\mu})'\mathbf{S}^{-1}(\overline{\mathbf{X}} - \boldsymbol{\mu}) \approx \chi^2_p$$

Large-Sample Behavior of $\overline{X}$ and S

Some additional considerations

- Using $\mathbf{S}$ instead of $\mathbf{\Sigma}$ does not seriously effect approximation.
- n must be large relative to p; that is, $(n-p)$ is large.
- The probability contours for $\overline{\mathbf{X}}$ are tighter than those for $\mathbf{X}$ since we have $(1/n)\mathbf{\Sigma}$ for $\overline{\mathbf{X}}$ rather than $\mathbf{\Sigma}$ for $\mathbf{X}$.

Comparison of Probability Contours

Below are contours for $99\%, 95\%, 90\%, 75\%, 50\%$ and 20% for an example with $n = 20$: