NMST432 ADVANCED REGRESSION MODELS

MAXIMUM LIKELIHOOD ESTIMATION THEORY
SUMMARY OF NOTATION AND MAIN RESULTS

Definition

Consider a random sample $\boldsymbol{X} = (X_1, \ldots, X_n)$ of independent identically distributed random variables (or vectors), each with density $f(x|\boldsymbol{\theta}_X)$ with respect to a σ-finite measure μ. We assume that $f(x|\boldsymbol{\theta}_X) \in \mathcal{F}$, where

$$\mathcal{F} = \{\text{distributions with density } f(x|\boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d\}$$

represents a parametric model for the distribution of the data.

The model $\mathcal{F}$ must satisfy the model identifiability condition: For any $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$ it holds $f(x|\boldsymbol{\theta}_1) \neq f(x|\boldsymbol{\theta}_2)$. In other words, no distribution can be parametrized by several different parameter vectors.

Because of independence, the joint denstity of the random sample $X_1, \ldots, X_n$ is $\prod_{i=1}^n f(x_i|\boldsymbol{\theta}_X)$. The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ of the parameter $\boldsymbol{\theta}_X$ is the point from Θ that maximizes the joint density evaluated at the observed values of $X_1, \ldots, X_n$.

Definition 1 (likelihood, log-likelihood)**.**

- The random function
$$L_n(\boldsymbol{\theta}) \stackrel{\mathsf{df}}{=} \prod_{i=1}^n f(X_i|\boldsymbol{\theta})$$
is called *the likelihood function* for the parameter $\boldsymbol{\theta}$ in the model $\mathcal{F}$.
- The random function
$$\ell_n(\boldsymbol{\theta}) \stackrel{\mathsf{df}}{=} \log L_n(\boldsymbol{\theta}) = \sum_{i=1}^n \log f(X_i|\boldsymbol{\theta})$$
is called *the log-likelihood function*.

Definition 2 (maximum likelihood estimator)**.** *The maximum likelihood estimator* (MLE) of the parameter $\boldsymbol{\theta}_X$ in the model $\mathcal{F}$ is defined as

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}\in\Theta} L_n(\boldsymbol{\theta}).$$

Note. Since the logarithm is strictly increasing, $L_n(\boldsymbol{\theta})$ and $\ell_n(\boldsymbol{\theta})$ attain the maximum at the same point.

Definition 3. Let P and Q be probability measures on the same probability space with densities p and q with respect to the same σ-finite measure μ (for example, $\mu = P + Q$). Define

$$K(P,Q) = \begin{cases} \mathsf{E}_P \log \frac{p(X)}{q(X)} = \int_{\{x:p(x)>0\}} \log \frac{p(x)}{q(x)} p(x)\, d\mu(x) & \text{if } \mathrm{P}\left[q(X) = 0\right] = 0 \\ +\infty & \text{otherwise.} \end{cases}$$

$K(P,Q)$ is called the *Kullback-Leibler distance (divergence)*.

Note. In fact, $K(P,Q)$ is a pseudo-distance: it holds $K(P,Q) \geq 0$, and $K(P,Q) = 0$ if and only if $P = Q$, but it is not symmetric: $K(P,Q) \neq K(Q,P)$.

Theorem 1. Suppose the support set $S = \{x \in \mathbb{R} : f(x|\boldsymbol{\theta}) > 0\}$ does not depend on the parameter $\boldsymbol{\theta}$. Denote P_X the induced probability measure of the random variable X_i and P_θ the probability measure associated with the density $f(x|\boldsymbol{\theta})$. Then for any $\boldsymbol{\theta} \neq \boldsymbol{\theta}_X$

$$\frac{1}{n}\log\frac{L_n(\boldsymbol{\theta}_X)}{L_n(\boldsymbol{\theta})} = \frac{1}{n}\sum_{i=1}^{n}\log\frac{f(X_i|\boldsymbol{\theta}_X)}{f(X_i|\boldsymbol{\theta})} \to K(P_X, P_\theta) \quad P_X - \text{ almost surely,}$$

and hence

$$\mathrm{P}\left[\ell_n(\boldsymbol{\theta}_X) > \ell_n(\boldsymbol{\theta})\right] \to 1 \quad \text{as } n \to \infty.$$

Note. When the number of observations increases to infinity, the (log-)likelihood function at the true parameter will be with a large probability larger than the (log-)likelihood function at any other parameter. This observation justifies the idea of estimating the parameters by maximizing the log-likelihood over all possible parameter vectors.

The calculation of the maximum likelihood estimator

The maximum likelihood estimator is usually determined by differentiation of the log-likelihood. The first derivative is set to zero and it is verified that the second derivative is negative definite.

Definition 4 (score, information)**.**

- The random vector
$$\boldsymbol{U}(\boldsymbol{\theta}|X_i) \overset{\mathrm{df}}{=} \frac{\partial}{\partial\boldsymbol{\theta}}\log f(X_i|\boldsymbol{\theta})$$
is called *the score function* for the parameter $\boldsymbol{\theta}$ in the model $\mathcal{F}$.
- The random vector
$$\boldsymbol{U}_n(\boldsymbol{\theta}|\boldsymbol{X}) \overset{\mathrm{df}}{=} \sum_{i=1}^{n}\boldsymbol{U}(\boldsymbol{\theta}|X_i) = \sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\log f(X_i|\boldsymbol{\theta})$$
is called *the score statistic*.
- The random matrix
$$I(\boldsymbol{\theta}|X_i) \overset{\mathrm{df}}{=} -\frac{\partial}{\partial\boldsymbol{\theta}^\mathsf{T}}\boldsymbol{U}(\boldsymbol{\theta}|X_i) = -\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^\mathsf{T}}\log f(X_i|\boldsymbol{\theta})$$
is called the contribution of the i-th observation to the information matrix.
- The random matrix
$$I_n(\boldsymbol{\theta}|\boldsymbol{X}) \overset{\mathrm{df}}{=} -\frac{1}{n}\frac{\partial}{\partial\boldsymbol{\theta}^\mathsf{T}}\boldsymbol{U}_n(\boldsymbol{\theta}|\boldsymbol{X}) = \frac{1}{n}\sum_{i=1}^{n}I(\boldsymbol{\theta}|X_i)$$
is called *the observed information matrix*.
- The matrix
$$I(\boldsymbol{\theta}) \overset{\mathrm{df}}{=} \mathsf{E}\,I(\boldsymbol{\theta}|X_i) = -\mathsf{E}\,\frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^\mathsf{T}}\log f(X_i|\boldsymbol{\theta})$$
is called *the expected (Fisher) information matrix*.

If the set Θ is open, the MLE $\widehat{\boldsymbol{\theta}}_n$ solves the system of equations $\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{X}) = \mathbf{0}$, that is

$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f(X_i|\widehat{\boldsymbol{\theta}}_n) = \mathbf{0}.$$

This system is called the *likelihood equations*.

The solution to the likelihood equations need not exist. Sometimes there may be multiple solutions, at most one of which is the MLE. If $I_n(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{X}) > 0$ (the observed information is positive definite at $\widehat{\boldsymbol{\theta}}_n$), we know that $\widehat{\boldsymbol{\theta}}_n$ is at least a local maximum. If $I_n(\boldsymbol{\theta}|\boldsymbol{X}) > 0$ for every $\boldsymbol{\theta} \in \Theta$, the log-likelihood function is concave and the solution to the likelihood equations must be the global maximum and hence the MLE.

In most cases no explicit solution can be found and the MLE must be calculated by numerical methods. There are two commonly used numerical methods for solving the likelihood equations. Let $\widehat{\boldsymbol{\theta}}^{(r)}$ be the r-th iteration to the solution.

- **The Newton-Raphson method:** $\widehat{\boldsymbol{\theta}}^{(r+1)} = \widehat{\boldsymbol{\theta}}^{(r)} + [nI_n(\widehat{\boldsymbol{\theta}}^{(r)}|\boldsymbol{X})]^{-1}\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}^{(r)}|\boldsymbol{X})$
- **The Fisher Scoring method:** $\widehat{\boldsymbol{\theta}}^{(r+1)} = \widehat{\boldsymbol{\theta}}^{(r)} + [nI(\widehat{\boldsymbol{\theta}}^{(r)})]^{-1}\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}^{(r)}|\boldsymbol{X})$

They are iterated until the change in $\widehat{\boldsymbol{\theta}}$ from one iteration to the next is sufficiently small or until $\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}})$ is sufficiently close to $\mathbf{0}$. The only difference between the two methods is in the information matrix: N-R uses the observed information, FS uses the expected information.

Both require setting $\widehat{\boldsymbol{\theta}}^{(1)}$, the starting value for numerical approximation, and are sensitive to its choice.

Properties of the maximum likelihood estimator

Maximum likelihood estimators are consistent and asymptotically normal as long as so called *regularity conditions* are satisfied.

Conditions (Regularity conditions for maximum likelihood estimators)**.**

R1. The number of parameters d in the model $\mathcal{F}$ is constant.

R2. The support set $S = \{x \in \mathbb{R} : f(x|\boldsymbol{\theta}) > 0\}$ does not depend on the parameter $\boldsymbol{\theta}$.

R3. The parameter space Θ is an open set.

R4. The density $f(x|\boldsymbol{\theta})$ is sufficiently smooth function of $\boldsymbol{\theta}$ (at least twice continuously differentiable).

R5. The Fisher information matrix $I(\boldsymbol{\theta})$ is finite, regular, and positive definite in a neighborhood of $\boldsymbol{\theta}_X$.

R6. The order of differentiation and integration can be interchanged in expressions such as

$$\frac{\partial}{\partial \boldsymbol{\theta}} \int h(x, \boldsymbol{\theta})\, d\mu(x) = \int \frac{\partial}{\partial \boldsymbol{\theta}} h(x, \boldsymbol{\theta})\, d\mu(x),$$

where $h(x, \boldsymbol{\theta})$ is either $f(x|\boldsymbol{\theta})$ or $\partial f(x|\boldsymbol{\theta})/\partial \boldsymbol{\theta}$.

Note. Take the identity

$$\int_{-\infty}^{\infty} f(x|\boldsymbol{\theta})\, d\mu(x) = 1$$

and differentiate both sides of the equation twice with respect to $\boldsymbol{\theta}$. Regularity condition R6 implies

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \boldsymbol{\theta}} f(x|\boldsymbol{\theta})\, d\mu(x) = \int_{-\infty}^{\infty} \frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathsf{T}}} f(x|\boldsymbol{\theta})\, d\mu(x) = \mathbf{0}. \tag{1}$$

Theorem 2 (consistency of the MLE)**.** Let conditions R1–R6 hold. Then there exists n_0 and a sequence $\widehat{\boldsymbol{\theta}}_n$ ($n \geq n_0$) of solutions to the likelihood equations $\boldsymbol{U}_n(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{X}) = \mathbf{0}$ such that $\widehat{\boldsymbol{\theta}}_n \xrightarrow{\mathsf{P}} \boldsymbol{\theta}_X$.

Note. If the log-likelihood is strictly concave, the likelihood equations have a unique solution, which is the MLE. It converges in probability to the true parameter. If the log-likelihood is not strictly concave, the likelihood equations may have multiple solutions representing local maxima and minima of the log-likelihood. There is one solution among them (the closest to $\boldsymbol{\theta}_X$), which provides a consistence sequence of estimators. Other solutions may not be close to $\boldsymbol{\theta}_X$ and may not converge to it.

Note. If there exists a sequence $\widetilde{\boldsymbol{\theta}}_n$ of other estimators that are guaranteed to be consistent (for example, moment estimators of $\boldsymbol{\theta}_X$), a consistent MLE can be obtained by taking the root of the likelihood equations, which is closest to $\widetilde{\boldsymbol{\theta}}_n$. Alternatively, one can perform one step of the Newton-Raphson algorithm with $\widetilde{\boldsymbol{\theta}}_n$ as the starting value.

Theorem 3 (Score function properties)**.** Let conditions R1–R6 hold. Then

(i) $\mathsf{E}\, \boldsymbol{U}(\boldsymbol{\theta}_X|X_i) = 0$, $\mathsf{var}\, \boldsymbol{U}(\boldsymbol{\theta}_X|X_i) = I(\boldsymbol{\theta}_X)$.

(ii) $\frac{1}{\sqrt{n}} \boldsymbol{U}_n(\boldsymbol{\theta}_X|\boldsymbol{X}) \xrightarrow{\mathsf{D}} \mathsf{N}_d(\mathbf{0}, I(\boldsymbol{\theta}_X))$.

Note. The Fisher information matrix at $\boldsymbol{\theta}_X$ can be calculated in two different ways: from Definition 4 (the expectation of minus the second derivative of the log density) or from Theorem 3 (the score function variance).

Theorem 4 (asymptotic normality of the MLE)**.** Suppose conditions R1–R6 hold. Let $\widehat{\boldsymbol{\theta}}_n$ be a consistent sequence of solutions to the likelihood equations. Then

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_d(\mathbf{0}, I^{-1}(\boldsymbol{\theta}_X)).$$

Note.

- The asymptotic variance of the MLE is equal to the inverse of the Fisher information. More information means better precision for estimation.
- The asymptotic variance of the MLE is in a certain sense optimal. Other estimators (e.g., moment estimators) cannot have a smaller asymptotic variance.

Theorem 5 (asymptotic distribution of the likelihood ratio)**.** Suppose conditions R1–R6 hold. Let $\widehat{\boldsymbol{\theta}}_n$ be a consistent sequence of solutions to the likelihood equations. Then

$$2 \log \frac{L_n(\widehat{\boldsymbol{\theta}}_n)}{L_n(\boldsymbol{\theta}_X)} = 2(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta}_X)) \xrightarrow{\mathsf{D}} \chi^2_d.$$

Theorem 6 (the Δ method for the MLE)**.** Suppose conditions R1–R6 hold. Let $\widehat{\boldsymbol{\theta}}_n$ be a consistent sequence of solutions to the likelihood equations. Take $q : \Theta \to \mathbb{R}^k$ a continuously differentiable function. Denote $\boldsymbol{\nu}_X = q(\boldsymbol{\theta}_X)$ a $D(\boldsymbol{\theta}) = \partial q(\boldsymbol{\theta})/\partial \boldsymbol{\theta}$. Then $\widehat{\boldsymbol{\nu}}_n = q(\widehat{\boldsymbol{\theta}}_n)$ is the MLE of the parameter $\boldsymbol{\nu}_X$ and

$$\sqrt{n}(\widehat{\boldsymbol{\nu}}_n - \boldsymbol{\nu}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_k(\mathbf{0}, D(\boldsymbol{\theta}_X) I^{-1}(\boldsymbol{\theta}_X) D(\boldsymbol{\theta}_X)^{\mathsf{T}}).$$

Tests based on maximum likelihood theory

The theory of the MLE can be used to derive tests of simple and composite hypotheses about the parameter $\boldsymbol{\theta}_X$.

Testing of simple hypotheses

We want to test the null hypothesis $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ against the alternative $H_1 : \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0 \in \Theta$. It is a simple hypothesis because there is just a single distribution in the model $\mathcal{F}$ with the density $f(x|\boldsymbol{\theta}_0)$.

We will introduce three different test statistics for testing H_0.

Definition 5.

(i) The statistic

$$\lambda_n = \frac{L_n(\widehat{\boldsymbol{\theta}}_n)}{L_n(\boldsymbol{\theta}_0)}$$

is called *the likelihood ratio*.

(ii) The statistic

$$W_n = n(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\mathsf{T} \widehat{I}_n(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

is called *the Wald statistic*.

(iii) The statistic

$$R_n = \frac{1}{n}\boldsymbol{U}_n(\boldsymbol{\theta}_0|\boldsymbol{X})^\mathsf{T} \widehat{I}_n^{-1}(\boldsymbol{\theta}_0)\boldsymbol{U}_n(\boldsymbol{\theta}_0|\boldsymbol{X})$$

is called *the Rao (score) statistic*.

Note. The symbol $\widehat{I}_n$ denotes any consistent estimator of the Fisher information matrix. Three different estimators can be used in Wald and Rao statistics:

1. $\widehat{I}_n(\boldsymbol{\theta}) = I_n(\boldsymbol{\theta}|\boldsymbol{X}) = -\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}^\mathsf{T}} \log f(\boldsymbol{\theta}|X_i)$ (the observed information matrix)
2. $\widehat{I}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{U}(\boldsymbol{\theta}|X_i)^{\otimes 2}$ (the empirical variance of the score function)
3. $\widehat{I}_n(\boldsymbol{\theta}) = I(\boldsymbol{\theta})$ (the Fisher information matrix)

The most common choice for the Wald statistic is $\widehat{I}_n(\widehat{\boldsymbol{\theta}}_n) = I_n(\widehat{\boldsymbol{\theta}}_n|\boldsymbol{X})$. The most common choice for the Rao statistic is $\widehat{I}_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{U}(\boldsymbol{\theta}_0|X_i)^{\otimes 2}$.

Note.

- The likelihood ratio requires the calculation of $\widehat{\boldsymbol{\theta}}_n$ and L_n or ℓ_n. It does not require the calculation of $\boldsymbol{U}_n$ and $\widehat{I}_n$.
- The Wald statistic requires the calculation of $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{I}_n$. It does not require the calculation of L_n and $\boldsymbol{U}_n$.
- Rao statistic requires the calculation of $\boldsymbol{U}_n$ a $\widehat{I}_n$. It does not require the calculation of $\widehat{\boldsymbol{\theta}}_n$ and L_n.

Note. If $d = 1$ (one parameter) and $\theta_0 = 0$, then the Wald statistic can be written as

$$W_n = \left[\frac{\widehat{\theta}_n}{\sqrt{n^{-1}\widehat{I}_n^{-1}(\widehat{\theta}_n)}} \right]^2,$$

where $n^{-1}\widehat{I}_n^{-1}(\widehat{\theta}_n)$ is the estimator of the asymptotic variance of $\widehat{\theta}_n$.

Theorem 7. Suppose conditions R1–R6 are satisfied. Let the hypothesis $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ hold. Then:

(i)
$$2\log\lambda_n = 2(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\boldsymbol{\theta}_0)) \xrightarrow{\mathsf{D}} \chi^2_d$$

(ii)
$$W_n \xrightarrow{\mathsf{D}} \chi^2_d$$

(iii)
$$R_n \xrightarrow{\mathsf{D}} \chi^2_d$$

Note. If H_0 holds, $\widehat{\boldsymbol{\theta}}_n$ should be close to $\boldsymbol{\theta}_0$, $L_n(\widehat{\boldsymbol{\theta}}_n)$ should be close to $L_n(\boldsymbol{\theta}_0)$, and $\boldsymbol{U}_n(\boldsymbol{\theta}_0|\boldsymbol{X})$ should be close to $\mathbf{0}$. Under H_0, all three test statistics have values close to 0. Their large values testify against H_0.

Corollary. Denote by $\chi^2_d(1-\alpha)$ the $(1-\alpha)$-quantile of χ^2_d distribution. Consider tests of $H_0 : \boldsymbol{\theta}_X = \boldsymbol{\theta}_0$ against $H_1 : \boldsymbol{\theta}_X \neq \boldsymbol{\theta}_0$ defined by the rule: reject H_0 in favor of H_1, if

(i) $2\log\lambda_n \geq \chi^2_d(1-\alpha)$ *(likelihood ratio test)*
(ii) $W_n \geq \chi^2_d(1-\alpha)$ *(Wald test)*
(iii) $R_n \geq \chi^2_d(1-\alpha)$ *(score test)*

Each of these tests has asymptotically (for $n \to \infty$) the level α.

Note. It can be shown that these three tests are asymptotically equivalent. For large sample sizes, their results are almost identical. With smaller sample sizes, their results can differ. Investigations of small sample behavior of these test statistics revealed that the likelihood ratio test has the best properties, the Wald test is the worst of the three.

Thus, in practical applications, the likelihood ratio test should be preferred.

Note. Under normality, the three test statistics are identical.

Estimation in the presence of nuisance parameters and testing of composite hypotheses

It is frequently desirable to estimate and test just a small number of parameters in a model that contains a much larger number of parameters. We divide the parameter vector into two subsets: the parameters of interest and the other paramaters – *nuisance parameters*.

Let $\boldsymbol{\theta}$ be divided into $\boldsymbol{\theta}_A$ containing the first m components of $\boldsymbol{\theta}$, and $\boldsymbol{\theta}_B$ containing the remaining $d - m$ components of $\boldsymbol{\theta}$. We have

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_A, \boldsymbol{\theta}_B)^\mathsf{T} = (\theta_1, \ldots, \theta_m, \theta_{m+1}, \ldots, \theta_d)^\mathsf{T}$$

We want to test the hypothesis $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$ against $H_1^* : \boldsymbol{\theta}_X \notin \Theta_0$, where $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0}\} \subset \Theta$. We want to know whether the first m components of $\boldsymbol{\theta}_X$ are equal to the vector of constants $\boldsymbol{\theta}_{A0}$ regardless of the other $d - m$ components of $\boldsymbol{\theta}_X$.

This is not a simple null hypothesis because there are many distributions in the model $\mathcal{F}$ that satisfy H_0^*.

All the vectors and matrices appearing in the notation of maximum likelihood estimation theory are decomposed into the first m components (part A) and the remaining $d - m$ components (part B). For example,

$$\widehat{\boldsymbol{\theta}}_n = \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{An} \\ \widehat{\boldsymbol{\theta}}_{Bn} \end{pmatrix}, \qquad \boldsymbol{U}_n(\boldsymbol{\theta}) = \begin{pmatrix} \boldsymbol{U}_{An}(\boldsymbol{\theta}) \\ \boldsymbol{U}_{Bn}(\boldsymbol{\theta}) \end{pmatrix}, \qquad I(\boldsymbol{\theta}) = \begin{pmatrix} I_{AA}(\boldsymbol{\theta}) & I_{AB}(\boldsymbol{\theta}) \\ I_{BA}(\boldsymbol{\theta}) & I_{BB}(\boldsymbol{\theta}) \end{pmatrix}, \quad \text{etc.}$$

The following lemma is useful for inverting the decomposed information matrix.

Lemma 8 (Block matrix inversion)**.** Let the matrix

$$I = \begin{pmatrix} I_{AA} & I_{AB} \\ I_{BA} & I_{BB} \end{pmatrix}$$

be of full rank. Then there exists an inverse matrix to I and it can be expressed as

$$I^{-1} = \begin{pmatrix} I^{AA} & I^{AB} \\ I^{BA} & I^{BB} \end{pmatrix},$$

where

$$\begin{aligned} I^{AA} &= I_{AA.B}^{-1}, \\ I^{AB} &= -I_{AA.B}^{-1} I_{AB} I_{BB}^{-1}, \\ I^{BA} &= -I_{BB.A}^{-1} I_{BA} I_{AA}^{-1}, \\ I^{BB} &= I_{BB.A}^{-1}, \\ I_{AA.B} &= I_{AA} - I_{AB} I_{BB}^{-1} I_{BA} \\ I_{BB.A} &= I_{BB} - I_{BA} I_{AA}^{-1} I_{AB}. \end{aligned}$$

If the null hypothesis $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$ holds we know that $\boldsymbol{\theta}_{AX} = \boldsymbol{\theta}_{A0}$, but we do not know the value of $\boldsymbol{\theta}_{BX}$. We can estimate $\boldsymbol{\theta}_{BX}$ by the maximum likelihood method applied to the nested submodel

$$\mathcal{F}_0 = \{\text{distributions with density } f(x|(\boldsymbol{\theta}_A, \boldsymbol{\theta}_B)),\ \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0},\ \boldsymbol{\theta}_B \in \Theta_B \subseteq \mathbb{R}^{d-m}\},$$

with $d - m$ unknown parameters.

Denote the maximum likelihood estimator of $\boldsymbol{\theta}_X$ in the submodel $\mathcal{F}_0$ by $\widetilde{\boldsymbol{\theta}}_n = \left(\begin{smallmatrix} \widetilde{\boldsymbol{\theta}}_{An} \\ \widetilde{\boldsymbol{\theta}}_{Bn} \end{smallmatrix}\right)$, where $\widetilde{\boldsymbol{\theta}}_{An} = \boldsymbol{\theta}_{A0}$ and $\widetilde{\boldsymbol{\theta}}_{Bn}$ solves the system of likelihood equations

$$\boldsymbol{U}_{Bn}(\boldsymbol{\theta}_{A0}, \widetilde{\boldsymbol{\theta}}_{Bn}) = \mathbf{0}.$$

The Fisher information matrix for $\boldsymbol{\theta}_B$ in this model is $I_{BB}(\boldsymbol{\theta}_X)$.

By Theorems 3 and 4 applied to the submodel $\mathcal{F}_0$, we get

$$\frac{1}{\sqrt{n}} \boldsymbol{U}_{Bn}(\boldsymbol{\theta}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_{d-m}(\mathbf{0}, I_{BB}(\boldsymbol{\theta}_X))$$

and

$$\sqrt{n}(\widetilde{\boldsymbol{\theta}}_{Bn} - \boldsymbol{\theta}_{BX}) \xrightarrow{\mathsf{D}} \mathsf{N}_{d-m}(\mathbf{0}, I_{BB}^{-1}(\boldsymbol{\theta}_X)).$$

On the other hand, Theorems 3 and 4 and Lemma 8 applied to the larger model imply

$$\frac{1}{\sqrt{n}}\boldsymbol{U}_{Bn}(\boldsymbol{\theta}_X) \xrightarrow{\mathsf{D}} \mathsf{N}_{d-m}(\mathbf{0}, I_{BB}(\boldsymbol{\theta}_X))$$

and

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_{Bn} - \boldsymbol{\theta}_{BX}) \xrightarrow{\mathsf{D}} \mathsf{N}_{d-m}(\mathbf{0}, I_{BB.A}^{-1}(\boldsymbol{\theta}_X)),$$

where (dropping the arguments $\boldsymbol{\theta}_X$)

$$I_{BB.A}^{-1} = (I_{BB} - I_{BA}I_{AA}^{-1}I_{AB})^{-1} \geq I_{BB}^{-1}.$$

Thus, the asymptotic variance of the MLE of the parameter $\boldsymbol{\theta}_{BX}$ depends on whether or not $\boldsymbol{\theta}_{AX}$ is known. If $\boldsymbol{\theta}_{AX}$ is known (which is true if H_0^* holds), the asymptotic variance of the MLE $\widetilde{\boldsymbol{\theta}}_{Bn}$ is generally larger than the asymptotic variance of the MLE $\widehat{\boldsymbol{\theta}}_{Bn}$ that does not assume a known $\boldsymbol{\theta}_{AX}$.

However, when $I_{BA} = 0$ (the estimators of $\boldsymbol{\theta}_{AX}$ and $\boldsymbol{\theta}_{BX}$ are asymptotically independent), then the asymptotic variances of $\widetilde{\boldsymbol{\theta}}_{Bn}$ and $\widehat{\boldsymbol{\theta}}_{Bn}$ are the same. Then it does not matter whether or not $\boldsymbol{\theta}_{AX}$ is known.

Let us generalize the three test statistics introduced in Definition 5 of the previous section to testing the composite hypothesis $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$ against $H_1^* : \boldsymbol{\theta}_X \notin \Theta_0$, where $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0}\} \subset \Theta$.

Definition 6.

(i) The statistic

$$\lambda_n^* = \frac{L_n(\widehat{\boldsymbol{\theta}}_n)}{L_n(\widetilde{\boldsymbol{\theta}}_n)}$$

is called *the likelihood ratio*.

(ii) The statistic

$$W_n^* = n(\widehat{\boldsymbol{\theta}}_{An} - \boldsymbol{\theta}_{A0})^\mathsf{T}\widehat{I}_{AA.B}(\widehat{\boldsymbol{\theta}}_n)(\widehat{\boldsymbol{\theta}}_{An} - \boldsymbol{\theta}_{A0})$$

is called *the Wald statistic*.

(iii) The statistic

$$R_n^* = \frac{1}{n}\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)^\mathsf{T}\widehat{I}_n^{-1}(\widetilde{\boldsymbol{\theta}}_n)\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)$$

is called *the Rao (score) statistic*.

Note.

- Obviously, $\lambda_n^* \geq 1$.
- The expression $\widehat{I}_{AA.B}$ in the Wald statistic means the inverse of the upper left block of the the matrix $\widehat{I}_n^{-1}$.
- Since $\boldsymbol{U}_{Bn}(\widetilde{\boldsymbol{\theta}}_n) = \mathbf{0}$, the Rao statistic can be written as

$$R_n^* = \frac{1}{n}\boldsymbol{U}_{An}(\widetilde{\boldsymbol{\theta}}_n)^\mathsf{T}\widehat{I}_{AA.B}^{-1}(\widetilde{\boldsymbol{\theta}}_n)\boldsymbol{U}_{An}(\widetilde{\boldsymbol{\theta}}_n).$$

- Theh Rao statistic does not require the calculation of the MLE $\widehat{\boldsymbol{\theta}}_n$ in the larger model, it only needs the MLE $\widetilde{\boldsymbol{\theta}}_n$ in the submodel. This is often much easier to get.

Theorem 9. Let the null hypothesis $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$, where $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0}\}$, hold. Then

(i)
$$2\log\lambda_n^* = 2(\ell_n(\widehat{\boldsymbol{\theta}}_n) - \ell_n(\widetilde{\boldsymbol{\theta}}_n)) \xrightarrow{\mathsf{D}} \chi_m^2;$$

(ii)
$$W_n^* \xrightarrow{\mathsf{D}} \chi_m^2;$$

(iii)
$$R_n^* \xrightarrow{\mathsf{D}} \chi_m^2.$$

Note. Under H_0^*, we expect $\widehat{\boldsymbol{\theta}}_n$ to be close to $\widetilde{\boldsymbol{\theta}}_n$, $L_n(\widehat{\boldsymbol{\theta}}_n)$ to be close to $L_n(\widetilde{\boldsymbol{\theta}}_n)$, and $\boldsymbol{U}_n(\widetilde{\boldsymbol{\theta}}_n)$ to be close to $\mathbf{0}$. The large values of the three test statistics testify against the null hypothesis.

Corollary. Let $\chi_m^2(1-\alpha)$ be $(1-\alpha)$-quantile of the χ_m^2 distribution. Consider tests of $H_0^* : \boldsymbol{\theta}_X \in \Theta_0$, where $\Theta_0 = \{\boldsymbol{\theta} : \boldsymbol{\theta}_A = \boldsymbol{\theta}_{A0}\}$, against $H_1^* : \boldsymbol{\theta}_X \notin \Theta_0$ given by the rule: reject H_0^* in favor of H_1^* if

(i) $2\log\lambda_n^* \geq \chi_m^2(1-\alpha)$ *(the likelihood ratio test)*
(ii) $W_n^* \geq \chi_m^2(1-\alpha)$ *(the Wald test)*
(iii) $R_n^* \geq \chi_m^2(1-\alpha)$ *(the score test)*

Then each of these three tests has asymptotically (for $n \to \infty$) the level α.

Note. The number of degrees of freedom in the reference χ_m^2 distribution is equal to the number of tested parameters.

Note. These three tests are asymptotically equivalent under the null hypothesis as well as under local alternatives. With small or moderate sample sizes, the likelihood ratio test has the best properties and the Wald test is the worst of the three. In practical applications, the likelihood ratio test should be preferred.

Note. Let $m = 1$, $\boldsymbol{\theta}_{AX} = \theta_{Xj}$, and $\boldsymbol{\theta}_{A0} = 0$. Consider the test of the hypothesis $H_0^* : \theta_{Xj} = 0$ against $H_1^* : \theta_{Xj} \neq 0$ (zero value of the j-th parameter in the presence of other parameters that are unspecified by the hypothesis). Then the Wald statistic can be written as

$$W_n = \left[\frac{\widehat{\theta}_{jn}}{\sqrt{n^{-1}\widehat{I}_{jj}^{-1}}}\right]^2,$$

where $n^{-1}\widehat{I}_{jj}^{-1}$ is the estimator of the asymptotic variance of $\widehat{\theta}_{jn}$. This is the square of the test statistic that statistical software typicaly evaluates to test zero value of a single model parameter.