

Lecture 11: Asymptotic normality and delta method

Lecturer: Yiqiao Zhong

Last update: 23:39:00, 2024-10-22

Disclaimer: materials in this pdf are used for STAT 709 “Mathematical Statistics” at the University of Wisconsin–Madison, but they are not guaranteed to be accurate. Consult textbooks whenever possible.

In classical asymptotic theory, one of the most useful results is asymptotic normality. Generally speaking, if θ is an unknown parameter, and $\hat{\theta}_n$ is a reasonable estimator (as usually stated across books/papers), then $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma_\theta^2)$ for certain variance σ_θ^2 . It is the basis for deriving p -values and confidence intervals for large-sample settings.

1 Convergence of transformations

First we introduce two useful theorems.

Theorem 1. (Continuous mapping theorem). Let $X, X_1, X_2, \dots$ be random vectors in $\mathbb{R}^d$ defined on a common probability space and g be a measurable function from $(\mathbb{R}^k, \mathcal{B}^k)$ to $(\mathbb{R}^l, \mathcal{B}^l)$. Suppose that g is continuous a.s. with respect to μ_X (namely the measure induced by X). Then

$$(i) \quad X_n \xrightarrow{a.s.} X \text{ implies } g(X_n) \xrightarrow{a.s.} g(X).$$

$$(ii) \quad X_n \xrightarrow{p} X \text{ implies } g(X_n) \xrightarrow{p} g(X).$$

$$(iii) \quad X_n \xrightarrow{d} X \text{ implies } g(X_n) \xrightarrow{d} g(X).$$

Proof. For (i), it is essentially pointwise convergence except for a zero-measure set. Here is a detailed proof. Let

$$A_0 = \{\omega \in \Omega : \lim_n X_n(\omega) = X(\omega)\}, \quad D = \{x \in \mathbb{R}^k : g \text{ is continuous at } x\}$$

By assumptions, $P(A_0^c) = 0$ and $\mu_X(D^c) = P(X^{-1}(D^c)) = 0$. Thus, the set $A = A_0 \cup X^{-1}(D)$ has probability measure 1. For every $\omega \in A$, we have pointwise convergence: $\lim_n X_n(\omega) = X(\omega)$ and g is continuous at $X(\omega)$, so $\lim_n g(X_n(\omega)) = g(X(\omega))$.

For (ii), our goal is to show, for any $\varepsilon > 0$, $\lim_n P(\|g(X_n) - g(X)\| > \varepsilon) = 0$. A proof is straightforward if we impose stronger assumption such as g is uniformly continuous or even lipschitz continuous, in which case $\|g(X_n) - g(X)\| \leq L\|X_n - X\|$ so

$$X_n \xrightarrow{p} X \implies \lim_n P(\|X_n - X\| > \varepsilon/L) = 0 \implies \lim_n P(\|g(X_n) - g(X)\| > \varepsilon) = 0.$$

The general case is proved in wiki for example.

For (iii), first we apply Skorohod's theorem (see [2][Thm. 3.2.8] alternatively): If $X_n \xrightarrow{d} X$, then there are random vectors $Y, Y_1, Y_2, \dots$ defined on a common probability space such that $\mu_Y = \mu_X$, $\mu_{Y_n} = \mu_{X_n}$, $n = 1, 2, \dots$, and $Y_n \xrightarrow{a.s.} Y$. Then we use Theorem 1 (i) and complete the proof. $\square$

Theorem 2 (Slutsky's theorem; Thm. 1.11 of [5]). Let $X, X_1, X_2, \dots, Y_1, Y_2, \dots$ be random variables on a probability space. Suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, where c is a fixed real number. Then

$$(i) \quad X_n + Y_n \xrightarrow{d} X + c.$$

$$(ii) \quad Y_n X_n \xrightarrow{d} cX.$$

$$(iii) \quad X_n/Y_n \xrightarrow{d} X/c \text{ if } c \neq 0.$$

In statistics, Y_n is usually a sequence of consistent estimators, so this theorem gives a limiting characterization.

Recall from the previous lecture that $Y_n \xrightarrow{P} c$ is equivalent to $Y_n \xrightarrow{d} c$. So Theorem 2 establishes arithmetic properties about convergence in distribution. It is important that we assume Y_n converges to a constant; otherwise none of the results are true.

If we think of $Z_n = (X_n, Y_n) \in \mathbb{R}^2$ as a two-dimension random vector, then convergence of $(X_n)_{n \geq 1}$ and $(Y_n)_{n \geq 1}$ only gives us information about the *marginal distribution* of Z_n , so we don't know how X_n and Y_n are *coupled*. Consider an example of perfect negative correlation: $X_n = -Y_n$ and both converge in distribution to $G \sim N(0, 1)$. Then, $X_n + Y_n \equiv 0$ instead of $2G$.

2 Delta method

A common scenario in statistics is to build new estimators from existing ones, so it is useful to know how to transfer the asymptotic properties. For example, $\{X_n\}$ is usually a sequence of estimators, then what is the property of $g(X_n)$?

Theorem 3 (Delta method). *Let $X_1, X_2, \dots$ and Y be random vectors in $\mathbb{R}^k$ satisfying*

$$a_n(X_n - c) \xrightarrow{d} Y, \tag{1}$$

where $c \in \mathbb{R}^k$ and $\{a_n\}$ is a sequence of positive numbers with $\lim_{n \rightarrow \infty} a_n = \infty$. Let g be a function on $\mathbb{R}^k$ to $\mathbb{R}$. If g is differentiable at c , then

$$a_n[g(X_n) - g(c)] \xrightarrow{d} [\nabla g(c)]^\top Y,$$

where $\nabla g(x) \in \mathbb{R}^k$ denotes the gradient of g at x .

In the special case where g is a linear function, then the result follows from Cramér-Wold theorem from Lecture 10. In general, we want to argue that g can be linearized in a local neighborhood around c .

Proof. Let

$$Z_n = a_n[g(X_n) - g(c)] - a_n[g(c)]^\top (X_n - c).$$

It suffices to show $Z_n = o_p(1)$, since the conclusion will follow by combining $Z_n = o_p(1)$ and Eqn. 1 due to Slutsky's theorem.

The differentiability of g at c implies that for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that

$$|g(x) - g(c) - [\nabla g(c)]^\top (x - c)| \leq \epsilon \|x - c\| \tag{2}$$

whenever $\|x - c\| < \delta_\epsilon$. Let $\eta > 0$ be any fixed scalar. When $|Z_n| \geq \eta$, then at least one of the two events happen: $\|X_n - c\| \geq \delta_\epsilon$ and $a_n \epsilon \|X_n - c\| \geq \eta$; this is because $\|X_n - c\| < \delta_\epsilon$ implies $|Z_n| \leq a_n \epsilon \|X_n - c\|$ due to Eqn. 2.

By the union bound, namely $A_0 \subset A_1 \cup A_2 \Rightarrow \mathbb{P}(A_0) \leq \mathbb{P}(A_1) + \mathbb{P}(A_2)$, we obtain

$$\mathbb{P}(|Z_n| \geq \eta) \leq \underbrace{\mathbb{P}(\|X_n - c\| \geq \delta_\epsilon)}_{I_1} + \underbrace{\mathbb{P}(a_n \epsilon \|X_n - c\| \geq \eta/\epsilon)}_{I_2}.$$

By the assumption $\lim_{n \rightarrow \infty} a_n = \infty$ and Slutsky's theorem, we have $X_n \xrightarrow{p} c$. So as $n \rightarrow \infty$,

$$I_1 = \mathbb{P}(\|X_n - c\| \geq \delta_\epsilon) \rightarrow 0$$

$$I_2 = \mathbb{P}(a_n \|X_n - c\| \geq \eta/\epsilon) \rightarrow 0.$$

Thus, $\lim_{n \rightarrow \infty} \mathbb{P}(|Z_n| \geq \eta) = 0$, which completes the proof. $\square$

As a simple consequence, when the asymptotic distribution Y is a normal distribution, then a continuous transformation yields another normal distribution.

Corollary 1 (delta method for asymptotic normality). *Assume the weak convergence of Eqn. 1. If $Y \sim N(0, \Sigma)$, then*

$$a_n[g(X_n) - g(c)] \xrightarrow{d} \mathcal{N}(0, [\nabla g(c)]^\top \Sigma \nabla g(c)). \quad (3)$$

Sometimes the first-order Taylor expansion is not sufficient for our purposes. The following example is taken from [5][Example. 1.31].

Example 1 (Higher-order delta method). *Let $(X_n)_{n \geq 1}$ be a sequence of random variables satisfying $\sqrt{n}(X_n - c) \xrightarrow{d} N(0, 1)$. Consider the function $g(x) = x^2$.*

1. *If $c \neq 0$, then the (first-order) delta method gives $\sqrt{n}(X_n^2 - c^2) \xrightarrow{d} N(0, 4c^2)$.*
2. *If $c = 0$, then we cannot apply the (first-order) delta method. Instead, if we consider second-order Taylor expansion and follow the proof strategy ([exercise](#)) of Theorem 1, then $nX_n^2 \xrightarrow{d} \chi_1^2$, where $\chi_1^2 \stackrel{d}{=} [N(0, 1)]^2$ is the chi-square distribution.*

In case 2, if we insist on using $\sqrt{n}X_n^2$ for delta method, we get convergence to a degenerate distribution where variance is zero. This means that we don't get useful distribution information, which motivates us to zoom in and consider a larger factor n instead of $\sqrt{n}$.

Plug-in estimator. Suppose we find $T_n(X)$ such that $\sqrt{n}(T_n(X) - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. If σ^2 is unknown, and we find $\hat{\sigma}_n^2$ such that $\hat{\sigma}_n^2$ is consistent for σ^2 . Then by Slutsky's theorem,

$$\frac{\sqrt{n}(T_n(X) - \theta)}{\hat{\sigma}_n} \xrightarrow{d} \mathcal{N}(0, 1).$$

3 Prevalence of asymptotic normality

In classical statistics, we expect that many common estimators have the asymptotic normality property, which forms the core of large-sample theory. For an estimator problem, we want to construct a consistent estimator with asymptotic variance (e.g., the variance in Eqn. 3) as small as possible, so there is less uncertainty and more accuracy in estimation.

Asymptotic Variance. Let $(a_n)_{n \geq 1}$ be a positive sequence, and either $a_n \rightarrow \infty$ or $a_n \rightarrow a > 0$. Assume

$$a_n(T_n(X) - \theta) \xrightarrow{d} Y \quad \text{with } 0 < \mathbb{E}Y^2 < \infty.$$

- (i) The asymptotic variance of the estimator $T_n(X)$ is defined to be

$$\frac{\text{Var}(Y)}{a_n^2}.$$

- (ii) Let $T'_n(X)$ be another estimator. Asymptotic relative efficiency is defined to be the ratio between two asymptotic variances.
- (iii) One important fact is

$$\mathbb{E}Y^2 \leq \liminf_{n \rightarrow \infty} \mathbb{E}[a_n^2(T_n - \theta)^2] \quad (4)$$

In many cases, Y has a symmetric distribution, so the asymptotic variance is a lower bound on $\mathbb{E}[(T_n - \theta)^2]$ asymptotically. It is possible that we have a strict inequality in Eqn. 4 and a very gap even exists; see Example 2 below.

Equation 4 is due to the following fact: if $X_n \xrightarrow{d} X$, then for any continuous (but possibly unbounded) function $f : \mathbb{R} \rightarrow \mathbb{R}$ satisfying $\mathbb{E}[|f(X)|] < \infty$ we have $\mathbb{E}[|f(X)|] \leq \liminf_{n \rightarrow \infty} \mathbb{E}[|f(X_n)|]$. (exercise: consider $g_M = \min\{|f|, M\}$ and then let $M \rightarrow \infty$.)

Example 2. Consider i.i.d. Bernoulli random variables $X_1, \dots, X_n$ with $P(X_i = 1) = p$ where $p \in (0, 1)$ is unknown. Both MoM and MLE lead to the sample mean estimator for p : $\hat{p}(X) = \bar{X}$. Now suppose we want to estimate $\theta = p^{-1}$. Consider the plug-in estimator

$$\hat{\theta} = \frac{1}{\bar{X}}.$$

We apply the delta method with transformation $g(x) = 1/x$ to obtain

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \frac{1-p}{p^3}\right)$$

The asymptotic variance is still the usual $O(1/n)$, but $\mathbb{E}[\hat{\theta}] = \infty$, that is, $\hat{\theta}$ does not even have a finite first moment.

Remark 1. Intuitively, convergence in distribution does not control the tail behavior. That is why the variance of the asymptotic distribution is a lower bound of the limiting variance.

3.1 Method of moments

Recall that we want to estimate a quantity $\theta = \theta(P) \in \mathbb{R}^k$ of an unknown distribution (either parametric or nonparametric). Suppose its k -th moment exists, and denote $\mu_j = \mathbb{E}_{X \sim P} X^j$ for $j \leq k$. Also suppose that we can find a one-one mapping $g : \mathbb{R}^k \rightarrow \mathbb{R}^k : \mu := (\mu_1, \dots, \mu_k) = g(\theta_1, \dots, \theta_k)$ and denote $h = g^{-1}$.

We use i.i.d. observations $X_1, \dots, X_n \in \mathbb{R}^d$ to construct empirical moments

$$\hat{\mu}_j = \frac{1}{n} \sum_{i=1}^n X_i^j$$

If h^{-1} is continuous, then we have known that by SLLN and continuous mapping theorem¹

$$\hat{\theta} = h^{-1}(\hat{\mu}) \xrightarrow{a.s.} \theta, \quad \text{as } n \rightarrow \infty.$$

To find the asymptotic distribution of MoM estimator $\hat{\theta}$, we make further assumptions: g is differentiable at μ and $\mathbb{E}[|X_1|^{2k}] < \infty$. Then, by an application of the delta method, $\hat{\theta}$ is asymptotically normal, satisfying

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, [\nabla g(\mu)]^\top V_\mu \nabla g(\mu)),$$

where $V_\mu \in \mathbb{R}^{k \times k}$ is given by

$$[V_\mu]_{ij} = \mu_{i+j} - \mu_i \mu_j.$$

¹As before, we suppress the dependence on n , we should have written $\hat{\theta}_n$, but normally it is clear from context that we view $n \rightarrow \infty$.

3.2 Maximum likelihood estimation

Let $P_\theta \in \mathcal{P}$ be an unknown distribution from a parametric family where $\theta \in \mathbb{R}^k$. Suppose we have n i.i.d. observation $X = (X_1, \dots, X_n)$ from P_θ with a common pdf $f_\theta(x)$. We know that the log likelihood function $\ell_n(\theta; X) = \sum_{i=1}^n \log f_\theta(X_i)$. Suppose that ℓ is differentiable with respect to θ . Let the MLE be $\hat{\theta}$, which by definition satisfies

$$\nabla_\theta \ell_n(\hat{\theta}; X) = 0.$$

We have learned the fact (without a proof) that MLE is consistent under some generic conditions. Let us derive the asymptotic distribution of $\hat{\theta}$ as a physicist.

Denote $Z_i = \log f_\theta(X_i)$, so $\ell(\theta; X) = \sum_{i=1}^n Z_i$. If θ is the unknown parameter such that $X_i \sim f_\theta$, then by Lecture 4, we have $\mathbb{E}Z_i = 0$. So by CLT,

$$\frac{1}{\sqrt{n}} \nabla_\theta \ell_n(\theta; X) \xrightarrow{d} N(0, V)$$

where $V = \mathbb{E} [\nabla_\theta [\log f_\theta(X_i)] \nabla_\theta [\log f_\theta(X_i)]^\top] \in \mathbb{R}^{k \times k}$. Since $\hat{\theta}$ and θ are very close due to consistency, we do a Taylor expansion,

$$\begin{aligned} \nabla_\theta \ell_n(\hat{\theta}; X) - \nabla_\theta \ell_n(\theta; X) &\approx \nabla_\theta^2 \ell_n(\theta; X) (\hat{\theta} - \theta) \\ &= nV (\hat{\theta} - \theta) \end{aligned}$$

The approximation step is not rigorous because we ignored the high-order terms and both sides involve random variables. Moreover, by LLN

$$\frac{1}{n} \nabla_\theta^2 \ell_n(\theta; X) \approx \mathbb{E} [\nabla_\theta^2 \log f_\theta(X_i)] = -V$$

where the last step is due to integration by parts ([exercise](#)). Thus we derive

$$\sqrt{n}(\hat{\theta} - \theta) \approx -\frac{1}{\sqrt{n}} V^{-1} \nabla_\theta \ell_n(\theta; X) \xrightarrow{d} N(0, V^{-1} V V^{-1}) = N(0, V^{-1})$$

This derivation suggests that the MLE is not only root- n consistent, but also has an asymptotically normal distribution. Notably, the matrix V is known as the **Fisher information** of the parametric family, and the asymptotic (co-)variance is optimal among a wide range of estimators—which we will discuss more in unbiased estimation.

3.3 Generalized method of moments*

Recall that we introduced a way of generalizing moments. Suppose $X_1, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} P_\theta$, $X_i \in \mathbb{R}^p$, $\theta \in \Theta$, where Θ is compact for simplicity. Let $g : \mathbb{R}^{p+1} \rightarrow \mathbb{R}^m$ be a continuous function, such that

$$\mathbb{E}[g(X_i, \theta)] = 0.$$

Note that m is allowed to be larger than p . Let $W \in \mathbb{R}^{m \times m}$ be a positive definite matrix. Define generalized method of moments (GMM) estimator as

$$\hat{\theta} \in \arg \min_{\theta \in \Theta} \left(\frac{1}{n} \sum_{i=1}^n g(X_i, \theta) \right)^\top W \left(\frac{1}{n} \sum_{i=1}^n g(X_i, \theta) \right).$$

We also write $\hat{\theta}(W, X)$ or simply $\hat{\theta}(W)$ to emphasize the dependence on W .

Consistency of GMM estimators. Suppose that $X_1, \dots, X_n \stackrel{i.i.d.}{\sim} P_{\theta_0}$, $\theta_0 \in \Theta$,

(1) (Identifiability). $W\mathbb{E}[g(X_i, \theta)] = 0$, if and only if $\theta = \theta_0$.

(2) (Regularity). $\mathbb{E}[\sup_{\theta \in \Theta} \|g(X_i, \theta)\|] < \infty$.

Then for any fixed W ,

$$\hat{\theta}(W) \xrightarrow{P} \theta_0, \text{ as } n \rightarrow \infty.$$

Remark 2. Condition (1) is necessary, because we need to make sure population minimizer is unique. Otherwise, g may be just duplicates of a single moment equation.

Proof idea. Denote

$$Q_n(\theta) = \left(\frac{1}{n} \sum_{i=1}^n g(X_i, \theta) \right)^T W \left(\frac{1}{n} \sum_{i=1}^n g(X_i, \theta) \right),$$

$$Q(\theta) = (\mathbb{E}[g(X_i, \theta)])^T W (\mathbb{E}[g(X_i, \theta)]).$$

Use ULLN to show

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{P} 0.$$

The minimizer of $Q(\theta)$ is unique $\theta = \theta_0$. So,

$$\hat{\theta} \xrightarrow{P} \theta_0.$$

(Recall Homework 4 Problem 6). □

Remark 3. Consistency holds if W is replaced by $(W_n)_{n \geq 1}$ with

$$W_n \xrightarrow{P} W.$$

Optimal weighting and asymptotic efficiency. Suppose $g \in C^1(\mathbb{R}^p \times \Theta)$. Denote

$$G = \mathbb{E}[\nabla_{\theta} g(X_i, \theta_0)] \in \mathbb{R}^{m \times p}, \quad \Omega = \mathbb{E}[g(X_i, \theta_0)g(X_i, \theta_0)^T] \in \mathbb{R}^{m \times m}.$$

Asymptotic Normality. In addition to conditions (1) and (2), assume that

(3) $\mathbb{E}[\sup_{\theta \in \Theta} \|\nabla g(X_i, \theta)\|] < \infty$.

(4) $G^T W G$ is nonsingular, which is satisfied if G has full column rank and $W \succ 0$.

Then,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \underbrace{(G^T W G)^{-1} G^T W \Omega W^T G (G^T W^T G)^{-1}}_{V(W)}).$$

Optimal weighting. Suppose $\Omega \succ 0$. Then

$$V(\Omega^{-1}) \preceq V(W), \quad \forall W \succ 0.$$

Here, $A_1 \succ A_2$ means $A_1 - A_2$ is positive definite, and $A_1 \succeq A_2$ means $A_1 - A_2$ is positive semidefinite. Asymptotic efficiency. We can first derive consistent estimator W_n of Ω^{-1} , i.e.

$$W_n \xrightarrow{P} \Omega^{-1},$$

and estimate θ_0 by $\hat{\theta}(W_n)$.

Conclusion: $\hat{\theta}(W_n)$ has the smallest asymptotic variance among all possible choices of W . GMM enlarges the scope of MoM, improves the estimation efficiency of MoM and enjoys nice theoretical guarantee.

4 Limitations of asymptotic approach

The asymptotic approach is widely used in statistical inference. Its validity depends on having a large sample size n so that the distribution $T_n(X)$ is approximated by its asymptotic one.

4.1 Small sample size

In n is not sufficiently large, there is an advantage to doing higher-order Taylor expansion. Let $X_1, \dots, X_n$ be i.i.d. random variables and $\bar{X}$ be the sample mean. Recall that in the proof of CLT, we expand the ch.f. function of $\bar{X}$ to the second order. It is possible to consider a higher-order expansion, which is known as **Edgeworth expansions**. We state a result from [5][Thm. 1.16].

Let $h : \mathbb{R} \rightarrow \mathbb{R}$ be a smooth function satisfying $h(\mu) = 0$ where $\mu = \mathbb{E}X_1$. Suppose that $\sigma_h^2 = \text{Var}(X_1)[h'(\mu)]^2 > 0$. We know that applying Delta method yields $W_n := \sqrt{n}h(\bar{X})/\sigma_h \xrightarrow{d} N(0, 1)$. Then under some conditions,

$$\sup_{x \in \mathbb{R}} \left| F_{W_n}(x) - \Phi(x) - \sum_{j=1}^m \frac{p_j(x)\Phi'(x)}{n^{j/2}} \right| = o\left(\frac{1}{n^{m/2}}\right)$$

where Φ is the normal CDF and p_j is certain polynomial of degree at most $3j - 1$.

4.2 Increasing dimensions

So far, when we establish consistency and asymptotic normality, we assume that the dimensions d of observations are fixed. In modern applications, the dimensions are quite large, from hundreds to millions. It seems inappropriate to imagine that the sample size n goes to infinity while the dimension d is fixed.

There are mainly two approaches to address this issue:

- (i) nonasymptotic approach where we spell out the dependence on dimension;
- (ii) asymptotic approach where we assume an appropriate scaling, e.g., a common regime is proportional growth: $d = d_n$ and $\lim_{n \rightarrow \infty} d_n/n = c$ where c is a constant.

For the asymptotic approach, we need to be clear about what consistency means when the dimension is growing. We need to specify what measurement (e.g., MSE) we make to assess the error of two random vectors of growing dimensions.

Example 3 (Dimension troubles in mean estimation). Suppose that $X_1, X_2, \dots, X_n \stackrel{\text{i.i.d.}}{\sim} N(\mu, I_d)$. Our goal is to estimate the unknown mean vector μ . Imagine that a sequence of estimation problems of growing sample size and dimension under the proportional regime: $d = d_n$ and $\lim_{n \rightarrow \infty} d_n/n = c$, also $\mu_n \in \mathbb{R}^{d_n}$ is of growing dimension (we emphasize the dependence on n in the subscript). Assume that the norm $\rho = \|\mu_n\| > 0$ is fixed.

Let $\hat{\mu}_n = \bar{X}$ be the sample mean estimator for μ_n . It is clear that $\hat{\mu}_n \sim N(\mu_n, \frac{1}{n}I_d)$. To assess the quality of the estimator, we consider **cosine similarity**:

$$\text{sim}(\hat{\mu}_n, \mu_n) = \frac{\langle \hat{\mu}_n, \mu_n \rangle}{\|\hat{\mu}_n\|_2 \|\mu_n\|_2}.$$

A geometric interpretation of cosine similarity is the cosine of the angle between the two vectors. Let $g \in N(0, I_d)$ be a standard normal variable, then

$$\text{sim}(\hat{\mu}_n, \mu_n) = \frac{\|\mu_n\|}{\|\mu_n + \frac{1}{\sqrt{n}}g\|} + \frac{1}{\sqrt{n}} \frac{\langle \mu_n / \|\mu_n\|, g \rangle}{\|\mu_n + \frac{1}{\sqrt{n}}g\|}.$$

To simplify the right-hand side, observe

$$\frac{\|\mu_n + \frac{1}{\sqrt{n}}g\|^2}{\|\mu_n\|^2} = 1 + \frac{2}{\rho^2\sqrt{n}}\langle\mu_n, g\rangle + \frac{1}{n\rho^2}\|g\|^2$$

By LLN, $\|g\|^2 = (1 + o_p(1))d$. Also, $\langle\mu_n, g\rangle \sim N(0, \|\rho\|^2)$. So we have

$$\frac{2}{\rho^2\sqrt{n}}\langle\mu_n, g\rangle = O_P\left(\frac{1}{\sqrt{n}}\right), \quad \frac{1}{n\rho^2}\|g\|^2 \xrightarrow{p} \frac{c}{\rho^2}.$$

Our derivation leads to

$$\frac{\|\mu_n + \frac{1}{\sqrt{n}}g\|}{\|\mu_n\|} \xrightarrow{p} \left(1 + \frac{c}{\rho^2}\right)^{1/2} \quad \frac{\langle\mu_n/\|\mu_n\|, g\rangle}{\|\mu_n + \frac{1}{\sqrt{n}}g\|} = O_P(1),$$

and thus

$$\text{sim}(\hat{\mu}_n, \mu_n) \xrightarrow{p} \left(1 + \frac{c}{\rho^2}\right)^{-1/2}.$$

If we measure estimation consistency under the cosine similarity, then the sample mean estimator is consistent if and only if $c = 0$.

5 Further reading

The book [1][Section 5.3] has an extensive introduction to delta method and various statistical applications. Jun Shao's book [5][Section 1.5.3] also provides the theory for delta method.

For GMM, a good reference includes [4], [3] and Wikipedia.

References

- [1] Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC, 2015.
- [2] Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 5th edition, 2019.
- [3] A.R. Hall. *Generalized Method of Moments*. Advanced texts in econometrics. Oxford University Press, 2005.
- [4] Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982.
- [5] Jun Shao. *Mathematical Statistics*. Springer Science & Business Media, 2nd edition, 2013.