LECTURE 10 NOTES

This lecture is a brief introduction to the modern topics of robust and high-dimensional statistics. We refer to Maronna, Martin and Yohai (2006) for an accesible account of robust statistics. High-dimensional statistics is a nascent area, and unfortunately, there are few monographs on the subject.

1. Robust estimators. Thus far, except for a few brief interludes, we considered the asymptotic behavior of estimators assuming the parametric model is well-specified; i.e. the (unknown) distribution of the observations is part of the parametric model. However, in practice, we cannot verify the correctness of the model, which leads us to consider *robust estimators*: estimators that are robust to small deviations from the parametric model.

What exactly are “small to medium deviations”? A common interpretation is an δ-contamination model:

$$\mathbf{x}_i \overset{i.i.d.}{\sim} \begin{cases} F_\theta & \text{w.p. } 1-\delta, \\ P & \text{w.p. } \delta, \end{cases}$$

where F_θ is in the parametric model but P is not. That is, δ fraction of the observations $\{\mathbf{x}_i\}$ are *outliers*.

Example 1.1. *An investigator wishes to estimate the location parameter of a Gaussian location model from observations* $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(x-\theta)$. *Unfortunately, unbeknownst to the investigator, some of the observations are corrupted. Thus the true generative model is*

$$\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} \begin{cases} f(x-\theta) & w.p.\ 1-\delta, \\ P & w.p.\ \delta, \end{cases}$$

where P is the (unknown) distribution of the outliers and δ is the (expected) fraction of outliers in the sample. Since the investigator is not aware of the outliers, he or she assumes the Gaussian location model is correct and estimates θ by the MLE $\bar{\mathbf{x}}$. It is possible to show that the variance of $\bar{\mathbf{x}}$ is

$$\mathbf{var}_\mu\big[\bar{\mathbf{x}}\big] = \frac{1}{n}\mathbf{var}_\mu\big[\mathbf{x}_1\big] = \frac{1}{n}\big((1-\delta)b_f^2 + \delta b_p^2 + \delta(1-\delta)(a-\theta)^2\big),$$

where a is the expected value of P and b_f^2, b_p^2 are the variances of F and P. If P is the standard Cauchy distribution, the variance of $\bar{\mathbf{x}}$ is unbounded.

Example 1.1 shows that even an arbitrarily small fraction of outliers may be catastrophic for a non-robust estimator. The preceding example leads us to consider the *breakdown point* of an estimator. Intuitively, it is the largest fraction of outliers an estimator can tolerate.

DEFINITION 1.2. *The finite-sample breakdown point of an estimator $\delta : \mathcal{X}^n \to \Theta$ at $x \in \mathcal{X}^n$ is*

$$\delta_b(x) := \tfrac{1}{n} \max\bigl\{k \in [n] : \sup_{x'} \bigl\|\delta(x) - \delta\bigl(x'\bigr)\bigr\| < \infty\bigr\},$$

where x and $x' \in \mathcal{X}^n$ differ in at most k observations.

Often, the breakdown point does not depend on the (unperturbed) observations $x \in \mathcal{X}^n$, and we drop the dependence on x: $\delta_b = \delta_b(x)$.

EXAMPLE 1.3 (Example 1.1 continued). *The breakdown point of $\bar{\mathbf{x}}$ is zero: by perturbing just one observation to be arbitrarily large, $\bar{\mathbf{x}}$ can be made arbitrarily large.*

Another estimator of θ is the median $\mathrm{med}(\mathbf{x})$. *It is possible to show that the median remains bounded as long as we perturb no more than $\lfloor\frac{n-1}{2}\rfloor$ observations. Indeed, as long as we perturb less than $\frac{n}{2}$ observations, the median, being larger than at least half of the observations, is larger than one of the unperturbed observations. Similarly, it is smaller than one of the unperturbed observations. However, if we are allowed to perturb $\lfloor\frac{n}{2}\rfloor$ observations, the median is unbounded. Thus its breakdown point is $\frac{1}{n}\lfloor\frac{n-1}{2}\rfloor$.*

The price of robustness is (asymptotic) efficiency. To quantify the loss of efficiency, we consider the ARE of the median.

LEMMA 1.4. *Let $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} F$. As long as*

1. *F is differentiable at θ, where θ is the population median.*
2. *$f(\theta) \neq 0$, f is the density of F.*

the sample median is asymptotically normal:

$$\sqrt{n}(\mathrm{med}(\mathbf{x}) - \theta) \xrightarrow{d} \mathcal{N}\bigl(0, (2f(\theta))^{-2}\bigr).$$

PROOF. To avoid complications, we assume n is odd. Let $\mathbf{y}_i := \mathbf{1}\bigl\{\mathbf{x}_i \le \theta + \frac{t}{\sqrt{n}}\bigr\}$. It is easy to check that $\mathbf{y}_i \sim \mathrm{Ber}(p_n)$, where $p_n := F\bigl(\theta + \frac{t}{\sqrt{n}}\bigr)$. By the fact that

$$\bigl\{\mathrm{med}(\mathbf{x}) \le \theta + \tfrac{t}{\sqrt{n}}\bigr\} \iff \bigl\{\textstyle\sum_{i\in[n]} \mathbf{y}_i \ge \frac{n+1}{2}\bigr\},$$

we have

$$\begin{aligned}\mathbf{P}\big(\sqrt{n}(\mathrm{med}(\mathbf{x})-\theta)\le t\big) &= \mathbf{P}\big(\mathrm{med}(\mathbf{x})\le\theta+\tfrac{t}{\sqrt{n}}\big)\\ &= \mathbf{P}\big(\textstyle\sum_{i\in[n]}\mathbf{y}_i\ge\frac{n+1}{2}\big)\\ &= \mathbf{P}\left(\frac{\sum_{i\in[n]}\mathbf{y}_i-np_n}{\sqrt{np_n(1-p_n)}}\ge\frac{\frac{n+1}{2}-np_n}{\sqrt{np_n(1-p_n)}}\right).\end{aligned}$$

By Lindeberg's CLT, $\frac{\frac{n+1}{2}-np_n}{\sqrt{np_n(1-p_n)}}\xrightarrow{d}\mathcal{N}(0,1)$. Thus

$$\mathbf{P}\left(\frac{\sum_{i\in[n]}\mathbf{y}_i-np_n}{\sqrt{np_n(1-p_n)}}\ge\frac{\frac{n+1}{2}-np_n}{\sqrt{np_n(1-p_n)}}\right)=\mathbf{P}\left(\mathbf{z}\ge\frac{\frac{n+1}{2}-np_n}{\sqrt{np_n(1-p_n)}}\right)+o(1),$$

where $\mathbf{z}\sim\mathcal{N}(0,1)$. Since F is differentiable at θ, it is possible to show that

$$\frac{\frac{n+1}{2}-np_n}{\sqrt{np_n(1-p_n)}}\to-2f(\theta)t.$$

Indeed,

$$\frac{\frac{n+1}{2}-np_n}{\sqrt{np_n(1-p_n)}}=\frac{\sqrt{n}\left(\frac{1}{2}-p_n\right)}{\sqrt{p_n(1-p_n)}}+\frac{\frac{1}{2}}{\sqrt{np_n(1-p_n)}}.$$

We drop the second term because it is $O\big(n^{-\frac{1}{2}}\big)$. Turning to the first term,

$$\begin{aligned}\frac{n\left(\frac{1}{2}-p_n\right)}{\sqrt{np_n(1-p_n)}} &= \frac{t}{\sqrt{p_n(1-p_n)}}\frac{F(\theta)-F\big(\theta+\frac{\theta}{\sqrt{n}}\big)}{\frac{t}{\sqrt{n}}}\\ &\to(2t)(-f(\theta))\end{aligned}$$

as $n\to\infty$. Finally, by the continuity of the Gaussian CDF,

$$\mathbf{P}\big(\sqrt{n}(\mathrm{med}(\mathbf{x})-\theta)\le t\big)\to\mathbf{P}\big(\mathbf{z}>-2f(\theta)t\big).$$

We rearrange the right side to obtain the stated conclusion. □

If $\mathbf{x}_i\overset{i.i.d.}{\sim}\mathcal{N}(\mu,1)$, the ARE of the median is $\frac{1}{4\phi(0)^2}\approx0.64$. We see that the median trades-off efficiency for robustness.

Both the mean and median are special cases of M-estimators:

$$\hat{\theta}_M\in\arg\min_{\theta\in\Theta}\sum_{i\in[n]}\rho(\mathbf{x}_i-\theta),$$

for some loss function $\rho:\Theta\to\mathbf{R}$. In particular, the mean minimizes

$$\textstyle\sum_{i\in[n]}(\mathbf{x}_i-\theta)^2,$$

and the median minimizes

$$\textstyle\sum_{i\in[n]} |\mathbf{x}_i - \theta| \,.$$

By choosing the loss function carefully, it is possible to adjust the trade-off between efficiency and robustness. Huber et al. (1964) proposed a compromise between the mean and median by minimizing what is now called Huber's loss:

$$\rho_\epsilon(x) := \begin{cases} \frac{1}{2}x^2 & |x| \le \epsilon \\ \epsilon|x| - \frac{\epsilon^2}{2} & |x| > \epsilon \end{cases}$$

for some parameter $\epsilon > 0$. Near zero, Huber's loss behaves like the square loss. Far away from zero, it behaves like the absolute value loss. The parameter ϵ adjusts the mix between the square loss and absolute value loss.

Lemma 1.5. *Let $\mathbf{x}_i \overset{\text{i.i.d.}}{\sim} f(x-\theta)$, where f is an even function. The asymptotic distribution of Huber's estimator is*

$$\sqrt{n}\big(\hat{\theta}_\epsilon - \theta\big) \xrightarrow{d} \mathcal{N}\Big(0, \tfrac{b_\epsilon^2 + (1-a_\epsilon)\epsilon^2}{a_\epsilon^2}\Big),$$

where $a_\epsilon = \mathbf{P}_0\big(|\mathbf{x}_1| \le \epsilon\big)$ and $b_\epsilon^2 = \mathbf{E}_0\Big[\mathbf{x}_1^2 \mathbf{1}_{[-\epsilon,\epsilon]}(\mathbf{x}_1)\Big]$.

Proof sketch. It is possible to show that Huber's loss is continuously differentiable. Let

$$\psi_\epsilon(x) := \partial\rho_\epsilon(x) = \begin{cases} \epsilon & x > \epsilon \\ x & x \in [-\epsilon, \epsilon] \\ -\epsilon & x < -\epsilon. \end{cases}$$

By the optimality of $\hat{\theta}_\epsilon$,

$$0 = \textstyle\sum_{i\in[n]} \psi_\epsilon(\mathbf{x}_i - \hat{\theta}_\epsilon).$$

We turn to our favorite trick (and dropping the remainder term) to obtain

$$0 \approx \textstyle\sum_{i\in[n]} \psi_\epsilon(\mathbf{x}_i - \theta) + \sum_{i\in[n]} \psi_\epsilon'(\mathbf{x}_i - \theta)(\hat{\theta}_\epsilon - \theta).$$

Rearranging,

$$\sqrt{n}(\hat{\theta}_\epsilon - \theta) \approx -\frac{\frac{1}{\sqrt{n}}\sum_{i\in[n]} \psi_\epsilon(\mathbf{x}_i - \theta)}{\frac{1}{n}\sum_{i\in[n]} \psi_\epsilon'(\mathbf{x}_i - \theta)}.$$

Since $f(x-\theta)$ is even and $\psi_\epsilon(x-\theta)$ is odd, the expected value of the numerator vanishes. Its variance is

$$\begin{aligned}&\mathbf{E}_\theta\Big[\psi_\epsilon(\mathbf{x}_1-\theta)^2\Big]\\ &=\textstyle\int_{\mathbf{R}}\psi_\epsilon(x_1-\theta)^2 f(x_1-\theta)dx.\end{aligned}$$

By a change of variables ($u = x_1 - \theta$),

$$\begin{aligned}&=\int_{\mathbf{R}}\psi_\epsilon(u)^2 f(u)du\\ &=\epsilon^2\mathbf{P}_0(\mathbf{x}_1>\epsilon)+\mathbf{E}_0\Big[\mathbf{x}_1^2\mathbf{1}_{[-\epsilon,\epsilon]}(\mathbf{x}_1)\Big]+\epsilon^2\mathbf{P}_0(\mathbf{x}_1<-\epsilon)\\ &=b_\epsilon^2+\epsilon^2(1-a_\epsilon).\end{aligned}$$

where $a_\epsilon = \mathbf{P}_0\big(|\mathbf{x}_1|\le\epsilon\big)$ and $b_\epsilon^2 = \mathbf{E}_0\Big[\mathbf{x}_1^2\mathbf{1}_{[-\epsilon,\epsilon]}(\mathbf{x}_1)\Big]$. By the CLT, the numerator is asymptotically normal:

$$\tfrac{1}{\sqrt{n}}\textstyle\sum_{i\in[n]}\psi_\epsilon(\mathbf{x}_i-\theta)\xrightarrow{d}\mathcal{N}(0,b_\epsilon^2+\epsilon^2(1-a_\epsilon)).$$

By a similar argument, the expected value of the denominator is

$$\mathbf{E}_\theta\Big[\psi_\epsilon'(\mathbf{x}_1-\theta)^2\Big]=\mathbf{E}_0\big[\mathbf{1}_{[-\epsilon,\epsilon]}(\mathbf{x}_1)\big]=a_\epsilon.$$

By Slutsky's theorem,

$$\sqrt{n}(\hat\theta_\epsilon-\theta)\xrightarrow{d}\mathcal{N}\Big(0,\tfrac{b_\epsilon^2+(1-a_\epsilon)\epsilon^2}{a_\epsilon^2}\Big),$$

which is the desired conclusion. □

Figure 1 shows the ARE of Huber's estimator as the parameter ϵ varies. If ϵ is large, Huber's loss behaves like the square loss whose minimizer is $\bar{\mathbf{x}}$. On the other hand, if ϵ is small, Huber's loss behaves like the absolute value loss whose minimizer is the median.

2. Superefficiency. We wish to declare an asymptotically efficient estimator has the smallest possible asymptotic variance. Unfortunately, the claim is not true without further qualification. As Example 2.1 shows, it is possible to construct estimators whose asymptotic variance is smaller than the Fisher information.

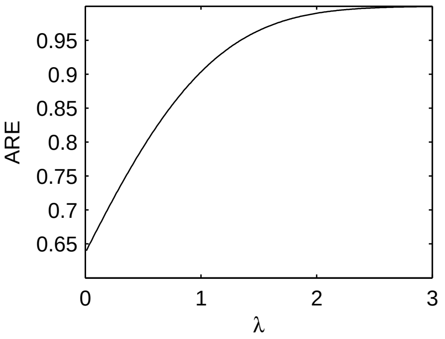

Fig 1: ARE of Huber's estimator ($\mathbf{x}_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$)

Example 2.1. *Let* $\{\mathbf{x}_i\}_{i\in[n]}$ *be* i.i.d. $\mathcal{N}(\mu,1)$ *random variables. Consider Hodges' estimator of* μ*:*

$$\hat{\mu}_n := \begin{cases} \bar{\mathbf{x}}_n & |\bar{\mathbf{x}}_n| > n^{-\frac{1}{4}}, \\ 0 & |\bar{\mathbf{x}}_n| \le n^{-\frac{1}{4}}. \end{cases}$$

Its distribution is a mixture of two components:

$$\begin{aligned} &\mathbf{P}_\mu(\sqrt{n}(\hat{\mu}_n - \mu) < t) \\ &\quad = \mathbf{P}_\mu(\sqrt{n}(\bar{\mathbf{x}}_n - \mu) < t)\mathbf{P}_\mu\big(|\bar{\mathbf{x}}_n| > n^{-\frac{1}{4}}\big) \\ &\qquad + \mathbf{1}_{[0,\infty)}(\sqrt{n}\mu + t)\mathbf{P}_\mu\big(|\bar{\mathbf{x}}_n| < n^{-\frac{1}{4}}\big). \end{aligned} \tag{2.1}$$

If $\mu \neq 0$*,*

$$\mathbf{P}_\mu\big(|\bar{\mathbf{x}}_n| > n^{-1/4}\big) = \mathbf{P}_\mu\big(\sqrt{n}\,|\bar{\mathbf{x}}_n| > n^{\frac{1}{4}}\big) = 1 - \Phi\big(n^{\frac{1}{4}} - \sqrt{n}\mu\big) \to 1.$$

Thus the first component dominates. Since $\xrightarrow{p}$ $\sqrt{n}(\bar{\mathbf{x}}_n - \mu) \xrightarrow{d} \mathcal{N}(0,1)$*,*

$$\mathbf{P}_\mu(\sqrt{n}(\hat{\mu}_n - \mu) < t) \to \Phi(t)$$

where $\Phi(t)$ *is the standard normal CDF. If* $\mu = 0$*,*

$$\mathbf{P}_\mu\big(|\bar{\mathbf{x}}_n| < n^{-\frac{1}{4}}\big) = \mathbf{P}_\mu\big(\sqrt{n}\,|\bar{\mathbf{x}}_n| < n^{\frac{1}{4}}\big) = \Phi\big(n^{\frac{1}{4}}\big) \to 1.$$

Thus the second component dominates, and we conclude the asymptotic distribution is degenerate:

$$\mathbf{P}_\mu(\sqrt{n}(\hat{\mu}_n - \mu) < t) \to \mathbf{1}_{[0,\infty)}(t)$$

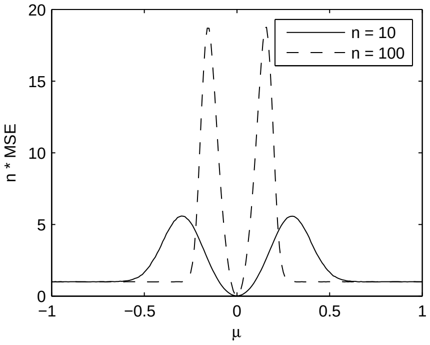

Fig 2: MSE of Hodges' estimator. The MSE of the MLE is constant at 1. Although Hodges estimator has smaller risk than the MLE near $\mu = 0$, its maximum risk is much larger.

We observe that the asymptotic variance of Hodges' estimator vanishes if $\mu = 0$ *and matches that of the MLE if* $\mu \neq 0$. *Thus it is asymptotically efficient at any* $\mu \neq 0$ *and superefficient at* $\mu = 0$.

What happened? In Hodges' example, the asymptotic picture is misleading: the convergence of the MSE to the asymptotic variance is not uniform in μ. Figure 2 shows the risk of Hodges' estimator: it fluctuates wildly near zero. The maximum risk of Hodges' estimator is also much larger than that of the MLE. The takeaway is point-wise asymptotics are misleading here.

More recently, rather than consider superefficiency a pathology, statisticians have sought to exploit superefficiency for statistical gain, especially on high-dimensional problems.

Example 2.2. *Consider the sparse denoising problem: estimate a sparse vector* $\mu \in \mathbf{R}^n$ *from observations* $\mathbf{x} \sim \mathcal{N}(\mu, I_n)$. *Such problems are common in signal processing, where signals are often sparse after transforming to an appropriate basis (e.g. wavelets).*

Since μ *is sparse, two intuitive estimators are the hard-thresholding estimator:*

$$\hat{\mu}_{\mathrm{HT}} := \begin{cases} \mathbf{x}_i & |\mathbf{x}_i| \geq \epsilon \\ 0 & |\mathbf{x}_i| < \epsilon, \end{cases}$$

and the soft-thresholding estimator:

$$\hat{\mu}_{\mathrm{ST},i} := \begin{cases} \mathbf{x}_i - \epsilon & \mathbf{x}_i \ge \epsilon \\ 0 & \mathbf{x}_i \in [-\epsilon, \epsilon] \\ \mathbf{x}_i + \epsilon & \mathbf{x}_i \le -\epsilon. \end{cases}$$

In this example, we focus on the soft-thresholding estimator since it generalizes easily to other problems. The analysis of the hard-threhsolding estimator is similar.

The question that remains is how to set the threshold ϵ. Intuitively, ϵ should be larger than the noise level, but not so large that it dominates the signal. At first, we may be inclined to set $\epsilon = 1$, but, as it turns out, $\epsilon = 1$ is too small. Since 1 is the average noise level; the noise may be much larger on some components, especially when n is large.

As it turns out, to account for the effect of n, an (almost) optimal choice is $\epsilon \approx \sqrt{2\log n}$. The choice is motivated by the fact that

$$\mathbf{E}\big[\max_{i\in[n]} |\mathbf{z}_i|\big] \sim O\big(\sqrt{2\log n}\big).$$

Indeed, for any $t > 0$, we have

$$\begin{aligned} &\exp\big(t\,\mathbf{E}\big[\max_{i\in[n]} |\mathbf{z}_i|\big]\big) \\ &\quad\le \mathbf{E}\big[\exp\big(t\max_{i\in[n]} |\mathbf{z}_i|\big)\big] && \textit{(Jensen's inequality)} \\ &\quad= \mathbf{E}\big[\max_{i\in[n]} \exp(t|\mathbf{z}_i|)\big] \\ &\quad\le \mathbf{E}\big[\textstyle\sum_{i\in[n]} \exp(t|\mathbf{z}_i|)\big]. \end{aligned}$$

Since $e^{|x|} \le e^x + e^{-x}$,

$$\begin{aligned} &\le \textstyle\sum_{i\in[n]} \mathbf{E}\big[\exp(t\mathbf{z}_i)\big] + \mathbf{E}\big[\exp(-t\mathbf{z}_i)\big] \\ &= n\big(\mathbf{E}\big[\exp(t\mathbf{z}_1)\big] + \mathbf{E}\big[\exp(-t\mathbf{z}_1)\big]\big) \\ &= 2n\exp\big(\tfrac{t^2}{2}\big). \end{aligned}$$

We plug in $t = \sqrt{2\log n}$ and rearrange to deduce

$$\mathbf{E}\big[\max_{i\in[n]} |\mathbf{z}_i|\big] \le \sqrt{2\log n} + \tfrac{\log 2}{\sqrt{2\log n}}.$$

It is possible to show a matching lower bound. By setting ϵ slightly larger than $\sqrt{2\log n}$, soft-thresholding correctly estimates all the components of μ that vanish.

We evaluate the MSE of the soft-thresholding estimator:

$$\mathbf{E}_\mu\Big[\|\hat{\mu}_{\mathrm{ST}} - \mu\|_2^2\Big] = \textstyle\sum_{i:\mu_i=0} \mathbf{E}_\mu\Big[(\hat{\mu}_{\mathrm{ST},i} - \mu_i)^2\Big] + \textstyle\sum_{i:\mu_i\neq 0} \mathbf{E}_\mu\Big[(\hat{\mu}_{\mathrm{ST},i} - \mu_i)^2\Big].$$

Since $\epsilon \approx \sqrt{2\log n}$, $\hat{\mu}_{\mathrm{ST},i} = 0$ for any $i \in [n] : \mu_i = 0$. Thus the first term is negligible. It is possible to show that

$$\textstyle\sum_{i:\mu_i\neq 0} \mathbf{E}_\mu\Big[(\hat{\mu}_{\mathrm{ST},i} - \mu_i)^2\Big] \sim O(s\log n),$$

where s is the sparsity of μ. Putting the pieces together, we deduce

$$\mathbf{E}_\mu\Big[\|\hat{\mu}_{\mathrm{ST}} - \mu\|_2^2\Big] \sim O(s\log n).$$

To evaluate the performance of soft-thresholding, consider the so-called oracle estimator that knows the support of μ. Since it knows which components are non-zero, it only has to estimate the s (non-zero) components, Thus the MSE of the oracle estimator is $O(s)$, which is only smaller than the MSE of $\hat{\mu}_{\mathrm{ST}}$ by a logarithmic factor! In other words, even if we knew the support of μ, we are only be able to improve upon the MSE of $\hat{\mu}_{\mathrm{ST}}$ by a small factor.

APPENDIX A: LINDEBERG'S CENTRAL LIMIT THEOREM

The phenomenon of asymptotic normality is not limited to sums of *i.i.d.* random variables. A more general setting is *triangular arrays.* A triangular array is a set of random variables arranged in a triangular array:

$$\begin{array}{lllll}
\mathbf{x}_{1,1} & & & & \\
\mathbf{x}_{2,1} & \mathbf{x}_{2,2} & & & \\
\vdots & \vdots & \ddots & & \\
\mathbf{x}_{n,1} & \mathbf{x}_{n,2} & \dots & \mathbf{x}_{n,n} & \\
\vdots & \vdots & & & \ddots,
\end{array}$$

where

1. random variables in the same row are mutually independent,
2. $\mathbf{E}\big[\mathbf{x}_{n,i}\big] = 0$ for any n, i,
3. $\sum_{i\in[n]} \mathbf{E}\big[\mathbf{x}_{i,j}^2\big] = 1$ for any n.

The third condition requires the triangular array to be correctly normalized. If $\mathbf{x}_i \overset{i.i.d.}{\sim} F$, where $\mathbf{E}\big[\mathbf{x}_i\big] = 0$ and $\mathbf{var}\big[\mathbf{x}_i\big] = 1$, letting $\mathbf{x}_{n,i} = \frac{1}{\sqrt{n}}\mathbf{x}_i$ ensures

$$\textstyle\sum_{i\in[n]} \mathbf{E}\big[\mathbf{x}_{n,i}^2\big] = n\,\mathbf{E}\big[\mathbf{x}_{n,1}^2\big] = \mathbf{E}\big[\mathbf{x}_1^2\big] = 1.$$

THEOREM A.1. *Let $\{\mathbf{x}_{i,j}\}$ be a triangular array. If the triangular array satisfies Lindeberg's condition*

$$\lim_{i\to\infty} \textstyle\sum_{i\in[n]} \mathbf{E}\big[\mathbf{x}_{n,i}^2 \mathbf{1}\{|\mathbf{x}_{n,i}| > \epsilon\}\big] = 0 \text{ for any } \epsilon > 0, \tag{A.1}$$

then $\sum_{i\in[n]} \mathbf{x}_{n,i}$ is asymptotically normal: $\sum_{i\in[n]} \mathbf{x}_{n,i} \overset{d}{\to} \mathcal{N}(0,1)$.

Theorem A.1 is called Lindeberg's CLT. Lindeberg's condition (A.1) essentially requires the variance of any $\mathbf{x}_{n,i}$ to vanish:

$$\begin{aligned}\mathbf{E}\big[\mathbf{x}_{n,i}^2\big] &= \mathbf{E}\big[\mathbf{x}_{n,i}^2 \mathbf{1}\{|\mathbf{x}_{n,i}| < \epsilon\}\big] + \mathbf{E}\big[\mathbf{x}_{n,i}^2 \mathbf{1}\{|\mathbf{x}_{n,i}| > \epsilon\}\big] \\ &\le \epsilon^2 + \mathbf{E}\big[\mathbf{x}_{n,i}^2 \mathbf{1}\{|\mathbf{x}_{n,i}| > \epsilon\}\big]\end{aligned}$$

Thus

$$\begin{aligned}\max_{i\in[n]} \mathbf{E}\big[\mathbf{x}_{n,i}^2\big] &= \epsilon^2 + \max_{i\in[n]} \mathbf{E}\big[\mathbf{x}_{n,i}^2 \mathbf{1}\{|\mathbf{x}_{n,i}| > \epsilon\}\big] \\ &= \epsilon^2 + \sum_{i\in[n]} \mathbf{E}\big[\mathbf{x}_{n,i}^2 \mathbf{1}\{|\mathbf{x}_{n,i}| > \epsilon\}\big],\end{aligned}$$

which vanishes by Lindeberg's condition. Unfortunately, Lindeberg's condition is hard to verify. A condition that is usually easier to verify is Lyapunov's condition:

$$\lim_{i\to\infty} \textstyle\sum_{i\in[n]} \mathbf{E}\Big[|\mathbf{x}_{n,i}|^{2+\epsilon}\Big] = 0 \text{ for some } \epsilon > 0. \tag{A.2}$$

COROLLARY A.2. *Under the conditions of Theorem A.1, if the triangular array satisfies Lyapunov's condition (A.2), then $\sum_{i\in[n]} \mathbf{x}_{n,i} \overset{d}{\to} \mathcal{N}(0,1)$.*

PROOF. We check that Lyapunov's condition implies Lindeberg's condition. For any $\epsilon > 0$,

$$\begin{aligned}|\mathbf{x}_{n,i}|^2 \mathbf{1}\{|\mathbf{x}_{n,i}| > \epsilon\} &= \frac{|\mathbf{x}_{n,i}|^{2+\epsilon} \mathbf{1}\{|\mathbf{x}_{n,i}| > \epsilon\}}{|\mathbf{x}_{n,i}|^{\epsilon} \mathbf{1}\{|\mathbf{x}_{n,i}| > \epsilon\}} \\ &\le \frac{|\mathbf{x}_{n,i}|^{2+\epsilon} \mathbf{1}\{|\mathbf{x}_{n,i}| > \epsilon\}}{\epsilon^{\epsilon}}.\end{aligned}$$

By the linearity of expectation,

$$\begin{aligned}\lim_{n\to\infty} \sum_{i\in[n]} \mathbf{E}\big[|\mathbf{x}_{n,i}|^2 \mathbf{1}\{|\mathbf{x}_{n,i}| > \epsilon\}\big] &\le \lim_{n\to\infty} \frac{1}{\epsilon^{\epsilon}} \sum_{i\in[n]} \mathbf{E}\big[|\mathbf{x}_{n,i}|^{2+\epsilon} \mathbf{1}\{|\mathbf{x}_{n,i}| > \epsilon\}\big] \\ &\le \frac{1}{\epsilon^{\epsilon}} \lim_{n\to\infty} \sum_{i\in[n]} \mathbf{E}\big[|\mathbf{x}_{n,i}|^{2+\epsilon}\big] \\ &= 0,\end{aligned}$$

which is Lindeberg's condition. □

REFERENCES

HUBER, P. J. et al. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics* **35** 73–101.

MARONNA, R., MARTIN, D. and YOHAI, V. (2006). *Robust statistics*. John Wiley & Sons, Chichester. ISBN.

YUEKAI SUN
BERKELEY, CALIFORNIA
DECEMBER 1, 2015