Characteristic functions

Patrick Breheny

September 19

Introduction

- Our next few lectures will focus on transformations and their distributions
- This is of constant practical use in statistics, as many complex estimators can be written as functions of simpler statistics with known convergence properties
- Before we do that, our main goal for today is to introduce a very useful tool known as the characteristic function that in many cases, greatly simplifies proofs of convergence

Helly-Bray Theorem

- Previously, we discussed the general conditions in which convergence in distribution implies convergence in mean (the dominated convergence theorem)
- We're going to start today by taking another look at that question, and specifically, at the question of when this is an "if and only if" situation
- The main result is summarized in the following theorem, which we will state without proof:
- **Theorem (Helly-Bray):** $\mathbf{x}_n \xrightarrow{\mathrm{d}} \mathbf{x}$ if and only if $\mathbb{E}g(\mathbf{x}_n) \to \mathbb{E}g(\mathbf{x})$ for all continuous bounded functions $g : \mathbb{R}^d \to \mathbb{R}$.
- Remark: Traditionally, "Helly-Bray Theorem" refers only to the forward part of the theorem

Almost everywhere

- The theorem can be extended in a few ways
- First, it doesn't have to be continuous everywhere; it can have discontinuities so long as they happen with probability zero
- **Definition:** Let $C(g) = \{\mathbf{x} : g \text{ is continuous at } \mathbf{x}\}$ denote the continuity set of a function $g : \mathbb{R}^d \to \mathbb{R}$. Then g is said to be *continuous almost everywhere* if $\mathbb{P}\{\mathbf{x} \in C(g)\} = 1$.
- This idea of something happening "almost everywhere" (i.e., with probability 1) is common in statistics: for example, we might refer to convergence of $f_n(x) \to f(x)$ almost everywhere, or a function being differentiable almost everywhere

Helly-Bray theorem, version 2

- We can now offer an alternate version of the Helly-Bray theorem
- **Theorem:** $\mathbf{x}_n \xrightarrow{\mathrm{d}} \mathbf{x}$ if and only if $\mathbb{E}g(\mathbf{x}_n) \to \mathbb{E}g(\mathbf{x})$ for all bounded measurable functions $g : \mathbb{R}^d \to \mathbb{R}$ such that $\mathbb{P}\{\mathbf{x} \in C(g)\} = 1$.
- Note that the reverse direction of this proof now follows directly from the definition of convergence in distribution

Closed sets

- Another way the theorem can be extended is by bounding the domain of g, as opposed to the range
- The technical condition we require is called compactness, which has an extremely abstract definition, but reduces to a very simple idea in $\mathbb{R}^d$
- First, we need to define the concepts of closed and open sets:
  - $\mathbf{x}$ is a *limit point* of a set A if for all ϵ, $N_\epsilon(\mathbf{x})$ contains at least one point in A other than $\mathbf{x}$.
  - A set A is *closed* if it contains all its limit points.
  - A set A is *open* if, for all $\mathbf{x} \in A$, there exists $N_\epsilon(\mathbf{x}) \subset A$.
- For example, $\{x : 0 < x < 1\}$ is open (and not closed) and $\{x : 0 \leq x \leq 1\}$ is closed (and not open)

Compact sets

- Now, for the abstract topological definition of compact:
- **Definition:** A collection $\{G_\alpha\}$ of open sets is said to be an *open cover* of the set A if $A \subset \cup_\alpha G_\alpha$. A set A is said to be *compact* if every open cover of A contains a finite subcover.
- Fortunately, in $\mathbb{R}^d$, we have the much simpler result that a set A is compact if and only if A is closed and bounded (in the sense that there exist $L, U : L \prec \mathbf{x} \prec U$ for all $\mathbf{x} \in A$)
- For example, the set $\{\mathbf{x} : a_i \leq x_i \leq b_i \text{ for all } i\}$, where $a_i < b_i$, is compact
- Lastly, a function $g : \mathbb{R}^d \to \mathbb{R}$ has *compact support* if there exists a compact set C such that $g(\mathbf{x}) = 0$ for all $\mathbf{x} \notin C$

Why is compactness important?

Compactness is important because many important properties of continuous functions only hold when the domain is a compact set:

- g continuous $\implies$ g bounded
- g continuous $\implies$ there exist $\mathbf{a}, \mathbf{b} : g(\mathbf{a}) = \inf g(\mathbf{x})$, $g(\mathbf{b}) = \sup g(\mathbf{x})$ (extreme value theorem)
- g continuous $\implies$ g uniformly continuous

Portmanteau theorem

To conclude, let's combine these statements (this is usually called the Portmanteau theorem, and can include several more equivalence conditions)

Theorem (Portmanteau): Let $g : \mathbb{R}^d \to \mathbb{R}$. The following conditions are equivalent:

(a) $\mathbf{x}_n \xrightarrow{\text{d}} \mathbf{x}$.

(b) $\mathbb{E}g(\mathbf{x}_n) \to \mathbb{E}g(\mathbf{x})$ for all continuous functions g with compact support.

(c) $\mathbb{E}g(\mathbf{x}_n) \to \mathbb{E}g(\mathbf{x})$ for all continuous bounded functions g.

(d) $\mathbb{E}g(\mathbf{x}_n) \to \mathbb{E}g(\mathbf{x})$ for all bounded measurable functions g such that g is continuous almost everywhere.

Portmanteau vs DCT

- Let's compare the Portmanteau and Dominated Convergence Theorems in terms of what we can conclude about expected values if we know that $\mathbf{x}_n \xrightarrow{\text{d}} \mathbf{x}$
  - Portmanteau: $\mathbb{E}g(\mathbf{x}_n) \to \mathbb{E}g(\mathbf{x})$ for g continuous, bounded
  - DCT: If $\|\mathbf{x}_n\| \leq Z$ and $\mathbb{E}Z < \infty$, then $\mathbb{E}\mathbf{x}_n \to \mathbb{E}\mathbf{x}$
- Students often ask whether one of these theorems is just a consequence of the other – the answer is no, they each say something different:
  - Portmanteau: Applies to any continuous function, but it has to be bounded
  - DCT: Applies only to $g(\mathbf{x}) = \mathbf{x}$, but works in unbounded cases

Characteristic functions

- In other words, with some qualifications, the argument that "moments converge, so distributions converge" is valid
- This fact is important for, say, moment generating functions; however, moment generating functions are unsatisfying because they do not always exist
- **Definition:** The *characteristic function* of a random variable $\mathbf{x} \in \mathbb{R}^d$ is $\varphi(\mathbf{t}) = \mathbb{E} \exp(i\mathbf{t}^\top \mathbf{x})$, where $i = \sqrt{-1}$.
- Remark: The characteristic function is the Fourier transform of the probability density (if you know what that is)

Continuity theorem

- We'll list helpful properties of characteristic functions in a moment, but let's begin by recognizing two critical ones:
    - For any random vector $\mathbf{x}$, $\varphi(\mathbf{t})$ exists and is continuous for all $\mathbf{t} \in \mathbb{R}^d$
    - Two random vectors $\mathbf{x}$ and $\mathbf{y}$ have the same distribution if and only if $\varphi_{\mathbf{x}}(\mathbf{t}) = \varphi_{\mathbf{y}}(\mathbf{t})$
- Furthermore, since $\exp(i\mathbf{t}^\top\mathbf{x}) = \cos(\mathbf{t}^\top\mathbf{x}) + i\sin(\mathbf{t}^\top\mathbf{x})$, we can immediately see the forward half of the following theorem (the other direction is much longer, so we're skipping it)
- **Theorem (Continuity):** $\mathbf{x}_n \xrightarrow{\text{d}} \mathbf{x}$ if and only if $\varphi_n(\mathbf{t}) \to \varphi(\mathbf{t})$ for all $\mathbf{t} \in \mathbb{R}^d$.

Properties of characteristic functions

We will now list, without proof, a bunch of helpful properties of characteristic functions ($b, \mathbf{c}$ constants, $\boldsymbol{\mu} = \mathbb{E}\mathbf{X}$)

(1) $\varphi(\mathbf{0}) = 1$ and $|\varphi(\mathbf{t})| \leq 1$ for all $\mathbf{t}$

(2) $\varphi_{\mathbf{x}/b}(\mathbf{t}) = \varphi_{\mathbf{x}}(\mathbf{t}/b)$ for $b \neq 0$

(3) $\varphi_{\mathbf{x}+\mathbf{c}}(\mathbf{t}) = \exp(i\mathbf{t}^\top \mathbf{c})\varphi_{\mathbf{x}}(\mathbf{t})$

(4) $\varphi_{\mathbf{x}+\mathbf{y}}(\mathbf{t}) = \varphi_{\mathbf{x}}(\mathbf{t})\varphi_{\mathbf{y}}(\mathbf{t})$ if $\mathbf{x} \perp\!\!\!\perp \mathbf{y}$

(5) $\nabla\varphi_{\mathbf{x}}(\mathbf{t})$ exists, is continuous, and $\nabla\varphi_{\mathbf{x}}(\mathbf{0}) = i\boldsymbol{\mu}$ if $\mathbb{E}\|\mathbf{x}\| < \infty$

(6) $\nabla^2\varphi_{\mathbf{x}}(\mathbf{t})$ exists, is continuous, and $\nabla^2\varphi_{\mathbf{x}}(\mathbf{0}) = -\mathbb{E}(\mathbf{x}\mathbf{x}^\top)$ if $\mathbb{E}\|\mathbf{x}\|^2 < \infty$

(7) $\varphi_{\mathbf{x}}(\mathbf{t}) = \exp(i\mathbf{t}^\top \mathbf{c})$ if $\mathbf{x} = \mathbf{c}$ with probability 1

(8) $\varphi_{\mathbf{x}}(\mathbf{t}) = \exp(i\mathbf{t}^\top \boldsymbol{\mu} - \frac{1}{2}\mathbf{t}^\top \boldsymbol{\Sigma}\mathbf{t})$ if $\mathbf{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Weak law of large numbers

- As mentioned, the main reason it helps to be familiar with characteristic functions is that they often provide a very convenient way to prove otherwise difficult theorems
- For example, let's return to the weak law of large numbers, which we stated without proof in the last set of notes
- **Theorem (Weak law of large numbers):** Let $\mathbf{x}, \mathbf{x}_1, \mathbf{x}_2, \ldots$ be independently and identically distributed random vectors such that $\mathbb{E}\|\mathbf{x}\| < \infty$. Then $\bar{\mathbf{x}}_n \xrightarrow{\text{P}} \boldsymbol{\mu}$, where $\boldsymbol{\mu} = \mathbb{E}(\mathbf{x})$.

Central limit theorem

- Similarly, proving the central limit theorem is equally straightforward
- **Theorem (Central limit):** Let $\mathbf{x}_1, \mathbf{x}_2, \ldots$ be iid random vectors with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. Then

$$\sqrt{n}(\bar{\mathbf{x}}_n - \boldsymbol{\mu}) \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma})$$

- In other words,
  - A first-order Taylor series expansion of the characteristic function gives us the WLLN
  - A second-order Taylor series expansion of the characteristic function gives us the CLT

Perhaps the two most important theorems in statistics, each with a simple four- or five-line proof!