STAT 383C: Statistical Modeling I **Fall 2016**

Lecture 1 — Aug 25

Lecturer: Purnamrita Sarkar *Scribe: Xueyu Mao, Jennifer Starling*

Disclaimer: These scribe notes have been slightly proofread and may have typos etc.

Note: The format of the scribe notes has been borrowed from EECS, U. C. Berkeley.

1.1 Different Types of Models

1.1.1 Parametric models

A parametric model is a set of distributions (or densities or regression functions) that can be parameterized by a finite number of parameters. For example, linear models.

Example 1.1. ***(Linear models).*** *Given inputs $x^{(1)}, \cdots, x^{(k)}$, the linear model to predict output y is:*

$$y = \sum_{i=1}^{k} \beta_i x^{(i)} + \varepsilon,$$

where ε is the intercept (Gaussian noise in some applications).

1.1.2 Nonparametric models

A nonparametric model is a set of distributions (or densities or regression functions) that can not be parameterized by a finite number of parameters.

Example 1.2. ***(Nonparametric density estimation).*** *$X_1, \cdots, X_n$ are observations from a cdf F, we want to estimate the pdf f, assuming some smoothness of f that $f \in \mathcal{F}_{DENS} \cap \mathcal{F}_{SOB}$, where $\mathcal{F}_{DENS}$ is a set of all probability density functions, and*

$$\mathcal{F}_{SOB} = \left\{ f : \int \left(f''(x) \right)^2 dx < \infty \right\}.$$

The class $\mathcal{F}_{SOB}$ is called a ***Sobolev space****.*

Example 1.3. ***(Regression).*** *Estimate $h(x) = \mathrm{E}(Y|X = x)$ using k observed pairs of data $(X_i, Y_i),\ i = 1, \cdots, k$. Regression model is like*

$$\mathcal{L} = h(x_1, \cdots, x_k) + \varepsilon,\ \varepsilon \sim Gaussian^1,$$

where $h(\cdot)$ is not necessarily linear and it is from a non-finite dimensional set.

[1] Note that while $\mathrm{E}[\varepsilon] = 0$ must hold, ε does not need to be Gaussian.

1.1.3 Semiparametric models

A semi-parmetric model can be partly written as parametric model and partly as nonparametric model. An example is like

$$\mathcal{L} = \beta_1 x_1 + \beta_2 x_2 + h(x_3, x_4, x_5) + \varepsilon,$$

which is a summation of linear model and regression model.

1.2 Convergence of Random Variables

Definition 1.1. *Let $X_1, X_2, \cdots$ be a sequence of random variables and let X be another random variable. We have the following two types of convergence:*

*1). X_n converges to X **in probability**, written $X_n \xrightarrow{P} X$, if $\forall \varepsilon > 0$,*

$$\mathrm{P}(|X_n - X| > \varepsilon) \to 0, \tag{1.1}$$

as $n \to \infty$.

*2). X_n converges to X **in distribution**, written $X_n \xrightarrow{d} X$ or $X_n \rightsquigarrow X$, if*

$$\lim_{n\to\infty} \mathrm{P}(X_n \le t) = \mathrm{P}(X \le t), \tag{1.2}$$

$\forall t$ s.t. $P(X \le t)$ is continuous at t.

Example 1.4. ***(Convergence in probability).*** *Suppose $X_n \sim \mathcal{N}(\mu, \sigma)$, then*

$$\mathrm{P}(\left|\overline{X_n} - \mu\right| > \varepsilon) = \mathrm{P}((\overline{X_n} - \mu)^2 > \varepsilon^2) \le \frac{\mathrm{E}\left[(\overline{X_n} - \mu)^2\right]}{\varepsilon^2} = \frac{Var(\overline{X_n})}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0,\ n \to 0,$$

where the "less than or equal to" is obtained by using Markov's inequality. You could also just use Chebyshev's inequality. So we have $\overline{X_n} \xrightarrow{P} \mu$. This is none other than our old friend – the Weak Law of Large Numbers.

Example 1.5. ***(Convergence in distribution).*** *Suppose $X_n \sim \mathcal{N}(\mu, \sigma)$, then*

$$\sqrt{n}(\overline{X_n} - \mu) \to N(0, 1)$$

This is none other than our old friend – the Central Limit Theorem.

Example 1.6. *Suppose $X_n \sim \mathcal{N}(0, \frac{1}{n})$, also by using Markov's inequality,*

$$\mathrm{P}(|X_n - 0| > \varepsilon) \le \frac{Var(X_n)}{\varepsilon^2} = \frac{1}{n\varepsilon^2} \to 0,$$

so $X_n \xrightarrow{P} 0$.

Now we have a point mass at 0 ($\mathrm{P}(X = 0) = 1$)*, then*

$$P(0 \le t) = \mathbf{1}(t \ge 0) \to \begin{cases} 1 & t \ge 0, \\ 0 & t < 0. \end{cases} \tag{1.3}$$

Remember that $\sqrt{n}X_n \sim \mathcal{N}(0,1)$*, consider*

$$\mathrm{P}(X_n \le t) = \mathrm{P}(\sqrt{n}X_n \le \sqrt{n}t) = \Phi(\sqrt{n}t) \to \begin{cases} 1 & t > 0, \\ \frac{1}{2} & t = 0, \\ 0 & t < 0. \end{cases} \tag{1.4}$$

Compare (1.3) and (1.4), we can conclude that $X_n \xrightarrow{d} 0$. *Note that although convergence fails at* $t = 0$*, the convergence in distribution also holds because the CDF of the limiting random variable* X *(which is 0 with probability 1) is not a continuous point at* 0 *(as shown in Figure 1.1).*

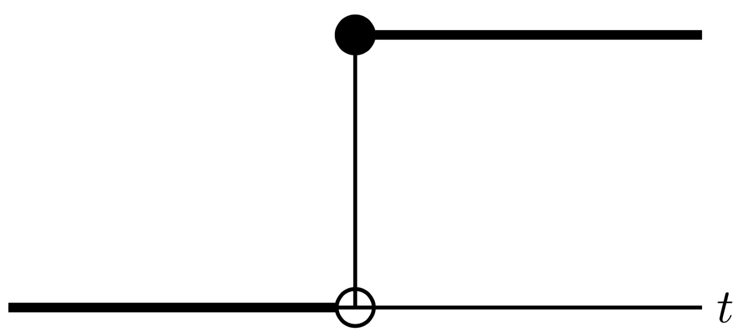

Figure 1.1. Jump function $\mathbf{1}(t \ge 0)$.

Remark 1.1. *We have the following relationship:*

$$X_n \xrightarrow{P} X \Rightarrow X_n \xrightarrow{d} X,$$

however, the reverse does not hold:

$$X_n \xrightarrow{d} X \nRightarrow X_n \xrightarrow{P} X.$$

Example 1.7. *A counter-example for the second part of Remark 1.1 is* $X \sim \mathcal{N}(0,1)$*, let* $X_n = -X \sim \mathcal{N}(0,1)$*, so* $X_n \xrightarrow{d} X$*. However,*

$$\mathrm{P}(|X_n - X| > \varepsilon) = \mathrm{P}(2\,|X_n| > \varepsilon) = \mathrm{P}(|X_n| > \frac{\varepsilon}{2}) \neq 0,$$

so $X_n \xrightarrow{P} X$ *does not hold.*

Remark 1.2. *X_n converges in probability to a random variable X does not imply the expectation of X_n converges to $E[X]$.*

$$X_n \xrightarrow{P} X,$$

$$\mathrm{E}[X_n] \nrightarrow E[X].$$

The point is that the tail is not well behaved. There is too much mass on the tail.

Example 1.8. *A counter-example for Remark 1.2 is*

$$X_n = \begin{cases} 0 & \text{with probability } 1 - \frac{1}{n}, \\ n^2 & \text{with probability } \frac{1}{n}, \end{cases}$$

We can see $X_n \xrightarrow{P} 0$ while $\mathrm{E}[X_n] = n \to \infty$.

Theorem 1.1. ***(Slutsky's theorem)*** *Given two sequences of random variables such that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$, then*

$$X_n + Y_n \xrightarrow{d} X + c,$$

and

$$X_n \cdot Y_n \xrightarrow{d} X \cdot c,$$

1.2.1 Delta Method

The Delta Method establishes the Convergence in Distribution of a transformation of a random variable under certain conditions.

Theorem 1.2. ***(The Delta Method)*** *Suppose that*

$$\frac{\sqrt{n}(Y_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

and that g is a differentiable function such that $g'(\mu) \neq 0$. Then

$$\frac{\sqrt{n}(g(Y_n) - g(\mu))}{|g'(\mu)|\sigma} \xrightarrow{d} N(0,1)$$

References

[1] Wasserman, Larry. All of statistics: a concise course in statistical inference. *Springer Science & Business Media*, 2013.