1 Basic concepts

1.1 Parametric vs Nonparametric models

A statistical model postulates a family of possible data generating mechanisms. Examples include:

(i) Let $X_1, \dots, X_n \sim^{\text{iid}} \Gamma(m, \theta)$ where m is known and $\theta \in (0, \infty) := \Theta$;

(ii) Let $Y_i = \alpha + \beta x_i + \varepsilon_i$ for $i \in [n] := \{1, \dots, n\}$, where $x_1, \dots, x_n$ and $\varepsilon_1, \dots, \varepsilon_n \sim^{\text{iid}} \mathcal{N}(0, \sigma^2)$. Here the unknown parameter is $\theta = (\alpha, \beta, \sigma^2) \in \mathbb{R} \times \mathbb{R} \times (0, \infty) := \Theta$.

If the parameter space Θ is finite-dimensional, we speak of a *parametric model*. When the model is correctly specified, i.e there exists $\theta_0 \in \Theta$ for which the data were generated from the distribution with parameter θ_0, typically we can use the MLE $\hat{\theta}_n$ to estimate θ_0, and expect $n^{1/2}(\hat{\theta}_n - \theta_0)$ to converge to a non-degenerate limiting distribution. On the other hand, when the model is misspecified, inferences may be very misleading.

Examples of nonparametric models include:

(i) Let $X_1, \dots, X_n \sim^{\text{iid}} F$ for some unknown distribution function F;

(ii) Let $X_1, \dots, X_n \sim^{\text{iid}} f$ for some density f belonging to some unknown smoothness class;

(iii) Let $Y_i = m(x_i) + \varepsilon_i$ for $i \in [n]$, where $x_1, \dots, x_n$ are known, m belongs to some unknown smoothness class and $\varepsilon_1, \dots, \varepsilon_n$ are iid with $\mathbb{E}(\varepsilon_1) = 0$, $\mathrm{Var}(\varepsilon_1) = \sigma^2$.

Such infinite-dimensional models are much less vulnerable to model misspecification. Typically however, we will pay a price in terms of a slower rate of convergence.

1.2 Estimating an arbitrary distribution function

Let $\mathcal{F}$ denote the set of all distribution functions on $\mathbb{R}$. The *empirical distribution function* $\mathbb{F}_n$ of real-valued random variables $X_1, \dots, X_n$ is defined by

$$\mathbb{F}_n(x) = \mathbb{F}_n(x, X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\{X_i \leq x\}.$$

Theorem (Glivenko-Cantelli Theorem). *Let $X_1, \dots, X_n \sim^{iid} F \in \mathcal{F}$ and let $\mathbb{F}_n$ denote the empirical distribution function of $X_1, \dots, X_n$. Then*

$$\sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \xrightarrow{a.s} 0 \text{ as } n \to \infty.$$

Proof. Let $\varepsilon > 0$ and $k := \lceil \frac{1}{\varepsilon} \rceil$. Let $x_0 = -\infty$, $x_i = \inf\{x \in \mathbb{R} : F(x) \geq i/k\}$ for $i \in [k-1]$ and $x_k = \infty$. Writing $F(x-)$ for $\lim_{y \uparrow x} F(y)$, note that for $i \in [k]$

$$F(x_i-) - F(x_{i-1}) \leq \frac{i}{k} - \frac{i-1}{k} = \frac{1}{k} \leq \varepsilon.$$

Now define the event

$$\Omega_{n,\varepsilon} = \left\{ \max_{i \in [k]} \sup_{m \geq n} |\mathbb{F}_m(x_i) - F(x_i)| \leq \varepsilon \right\} \cap \left\{ \max_{i \in [k]} \sup_{m \geq n} |\mathbb{F}_m(x_i-) - F(x_i-)| \leq \varepsilon \right\}$$

Noting that both $\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i \leq x\}$ and $\mathbb{F}_n(x-) = \frac{1}{n} \sum_{i=1}^n \mathbb{1}\{X_i < x\}$ are both sample averages of i.i.d random variables, we have by a union bound and the SLLN that

$$\begin{aligned}
&\mathbb{P}_F(\Omega_{n,\varepsilon}^c) \\
&\leq \sum_{i=1}^k \mathbb{P}_F \left(\sup_{m \geq n} |\mathbb{F}_m(x_i) - F(x_i)| > \varepsilon \right) + \sum_{i=1}^k \mathbb{P}_F \left(\sup_{m \geq n} |\mathbb{F}_m(x_i-) - F(x_i-)| > \varepsilon \right) \\
&\xrightarrow{a.s} 0.
\end{aligned}$$

□