1 Basic concepts

1.1 Parametric vs Nonparametric models

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

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

(ii) Let $Y_i = \alpha + \beta x_i + \varepsilon_i$ for $i \in [n] := \{1, \ldots, n\}$, where $x_1, \ldots, x_n$ and $\varepsilon_1, \ldots, \varepsilon_n \sim^{\text{iid}} \mathcal{N}(0, \sigma^2)$. 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 $X_1, \ldots, X_n \sim^{\text{iid}} F$ for some unknown distribution function F;

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

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

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 $X_1, \ldots, X_n$ is defined by

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

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

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

Now let $x \in \mathbb{R}$ and find $i_* \in [k]$ such that $x \in [x_{i_*-1}, x_{i_*})$. Then for any $n_0 \in \mathbb{N}$ and $n \geq n_0$,

$$\begin{aligned}
\mathbb{F}_n(x) - F(x) &\leq \mathbb{F}_n(x_{i_*}-) - F(x_{i_*-1}) \\
&= \mathbb{F}_n(x_{i_*}-) - F(x_{i_*}-) + F(x_{i_*}-) - F(x_{i_*-1}) \\
&\leq \max_{i \in [k]} \sup_{m \geq n_0} |\mathbb{F}_m(x_i-) - F(x_i-)| + \varepsilon.
\end{aligned}$$

We also have

$$\begin{aligned}
F(x) - \mathbb{F}_n(x) &\leq F(x_{i_*}-) - \mathbb{F}_n(x_{i_*-1}) \\
&= F(x_{i_*}-) - F(x_{i_*-1}) + F(x_{i_*-1}) - \mathbb{F}_n(x_{i_*-1}) \\
&\leq \varepsilon + \max_{i \in [k]} \sup_{m \geq n_0} |\mathbb{F}_m(x_i) - F(x_i)|.
\end{aligned}$$

It follows that

$$\mathbb{P}_F \left(\sup_{n \geq n_0} \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| > 2\varepsilon \right) \leq \mathbb{P}_F(\Omega_{n_0,\varepsilon}^c) \to 0 \text{ as } n_0 \to \infty.$$

Since $\varepsilon > 0$ was arbitrary, we conclude that

$$\begin{aligned}
\mathbb{P}_F \left(\sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \to 0 \right) &= \mathbb{P}_F \left(\bigcap_{L=1}^\infty \bigcup_{n_0=1}^\infty \left\{ \sup_{n \geq n_0} \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \leq \frac{1}{L} \right\} \right) \\
&= \lim_{L \to \infty} \lim_{n_0 \to \infty} \mathbb{P}_F \left(\sup_{n \geq n_0} \sup_{x \in \mathbb{R}} |\mathbb{F}_n(x) - F(x)| \leq \frac{1}{K} \right) \\
&= 1.
\end{aligned}$$

□

In fact, we can say much more:

Theorem (Dvoretzky-Kiefer-Wolfowitz Theorem)**.** *Under the conditions of the previous theorem, for every $\varepsilon > 0$,*

$$\mathbb{P}_F\left(\sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F(x)| > \varepsilon\right) \leq 2e^{-2n\varepsilon^2}.$$

Proof. Not given. □

Corollary (Uniform Glivenko-Cantelli)**.** *Under the conditions of the Glivenko-Cantelli Theorem,*

$$\sup_{F\in\mathcal{F}} \mathbb{P}_F\left(\sup_{m\geq n}\sup_{x\in\mathbb{R}} |\mathbb{F}_m(x) - F(x)| > \varepsilon\right) \to 0 \text{ as } n \to \infty.$$

Proof. By a union bound and the DKW inequality,

$$\begin{aligned}\sup_{F\in\mathcal{F}} \mathbb{P}_F\left(\sup_{m\geq n}\sup_{x\in\mathbb{R}} |\mathbb{F}_m(x) - F(x)| > \varepsilon\right) &\leq \sup_{F\in\mathcal{F}} \sum_{m\geq n} \mathbb{P}_F\left(\sup_{x\in\mathbb{R}} |\mathbb{F}_m(x) - F(x)| > \varepsilon\right) \\ &\leq 2\sum_{m\geq n} e^{-2m\varepsilon^2} \\ &\frac{2e^{-2n\varepsilon^2}}{1-e^{-2\varepsilon^2}} \to 0 \text{ as } n \to \infty.\end{aligned}$$

□

As another application of the DKW inequality, consider the problem of finding a confidence bound for F. Given $\alpha \in (0,1)$, set $\varepsilon_n = \varepsilon_n(\alpha) := \left\{\frac{1}{2n}\log\left(\frac{2}{\alpha}\right)\right\}^{1/2}$. Then, by the DKW Theorem,

$$\mathbb{P}_F\left(\max\{0, \mathbb{F}_n(x) - \varepsilon_n\} \leq F(x) \leq \min\{\mathbb{F}_n(x) + \varepsilon_n, 1\} \ \ \forall x \in \mathbb{R}\right) \geq 1 - \alpha.$$

In fact, let $U_1, \ldots, U_n \sim^{\text{iid}} \mathcal{U}[0,1]$ and let $\mathbb{G}_n$ denote their empirical distribution. Define the *quantile function* $F^{-1} : (0,1] \to (-\infty, \infty]$ by $F^{-1}(p) := \inf\{x \in \mathbb{R} : F(x) \geq p\}$ (i.e the *generalised inverse*). Since F is increasing and right-continuous, we have $\{x \in \mathbb{R} : F(x) \geq p\} = [F^{-1}(p), \infty)$, so $\{U_i \leq F(x)\} = \{F^{-1}(U_i) \leq x\}$. Hence

$$\begin{aligned}\mathbb{G}_n(F(x)) = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{U_i \leq F(x)\} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{F^{-1}(U_i) \leq x\} &=^d \frac{1}{n}\sum_{i=1}^n \mathbb{1}\{X_i \leq x\} \\ &= \mathbb{F}_n(x).\end{aligned}$$

So

$$\sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F(x)| =^d \sup_{x\in\mathbb{R}} |\mathbb{G}_n(F(x)) - F(x)| \leq \sup_{t\in[0,1]} |\mathbb{G}_n(t) - t|$$

with equality if F is continuous. Thus if F is continuous, the distribution of $\sup_{x\in\mathbb{R}} |\mathbb{F}_n(x) - F(x)|$ does not depend on F!

Other generalisations of the Glivenko-Cantelli Theorem include Uniform Laws of Large Numbes (ULLN). Let $X_1, X_2, \ldots, X_n$ be iid, taking values in some measurable space $(\mathcal{X}, \mathcal{A})$, and let $\mathcal{G}$ denote a family of real-valued measurable functions on $\mathcal{X}$. We say $\mathcal{G}$ satisfies a *ULLN* if

$$\sup_{g\in\mathcal{G}} \left| \frac{1}{n}\sum_{i=1}^{n} g(X_i) - \mathbb{E}g(X) \right| \xrightarrow{a.s} 0.$$

Remark. $\mathcal{G}$ may be uncountable, so this random variable could be non measurable. But we can get around this by working with an outer probability $\mathbb{P}^*(A) = \inf\{\mathbb{P}(B) : B \in \mathcal{F},\ B \supseteq A\}$.

Thus, in the Glivenko-Cantelli Theorem we proved that $\{\mathbb{1}\{\cdot \leq x\} : x \in \mathbb{R}\}$ satisfies a ULLN. In general, proving a ULLN requires control of the 'size' of $\mathcal{G}$, which can be measured for instance through its metric entropy (see van de Geer, 2000).

1.3 Concentration Inequalities

Definition. A random variable X is *sub-Gaussian with variance parameter* σ^2 if

$$\mathbb{E}(e^{tX}) \leq e^{t^2\sigma^2/2} \quad \forall t \in \mathbb{R}.$$

Remark. It can be shown (see Example Sheet) that any such random variable must satisfy $\mathbb{E}(X) = 0$ and $\mathrm{Var}(X) \leq \sigma^2$. Note also that we have equality in the definition if $X \sim \mathcal{N}(0, \sigma^2)$.

Here are some characterisations of sub-Gaussianity:

Proposition.

(a) If X is sub-Gaussian with variance parameter σ^2, then

$$\max\{\mathbb{P}(X \geq x), \mathbb{P}(X \leq -x)\} \leq e^{-\frac{x^2}{2\sigma^2}} \qquad (*)$$

for every $x \geq 0$.

(b) If X satisfies $(*)$, then for every $q \in \mathbb{N}$,

$$\mathbb{E}(X^{2q}) < 2q!(2\sigma^2)^q \leq q!(4\sigma^2)^q.$$

(c) Suppose that $\mathbb{E}(X) = 0$ and $\mathbb{E}(X^{2q}) \leq q!C^{2q}$ for every $q \in \mathbb{N}$. Then X is sub-Gaussian with variance parameter $4C^2$.

Proof.

(a) By Markov's inequality (i.e Chernoff's bound)

$$\mathbb{P}(X \geq x) \leq \inf_{t \geq 0} e^{-tX}\mathbb{E}(e^{tX}) \leq \inf_{t \geq 0} e^{-tx+t^2\sigma^2/2} = e^{-\frac{x^2}{2\sigma^2}}$$

for every $x \geq 0$, since the infimum is attained at $t = \frac{x}{\sigma^2}$. The bound for $\mathbb{P}(X \leq -x)$ is similar (since $-X$ is also σ^2-sub-Gaussian).

(b) We have

$$\begin{aligned}
\mathbb{E}(X^{2q}) = \int_0^\infty \mathbb{P}(X^{2q} \geq x)\mathrm{d}x &= 2q\int_0^\infty y^{2q-1}\mathbb{P}(|X| \geq y)\mathrm{d}x \\
&\leq 4q\int_0^\infty y^{2q-1}e^{-\frac{y^2}{2\sigma^2}}\mathrm{d}y \\
&= 2q(2\sigma^2)^q\int_0^\infty t^{q-1}e^{-t}\mathrm{d}t \qquad (t = \tfrac{y^2}{2\sigma^2}) \\
&= 2q!(2\sigma^2)^q
\end{aligned}$$

for $q \in \mathbb{N}$.

(c) Let X' denote an independent copy of X. Then $X - X'$ has a symmetric distribution, so for every $t \in \mathbb{R}$,

$$\begin{aligned}
\mathbb{E}(e^{t(X-X')}) &= \frac{1}{2}\left(\mathbb{E}(e^{t(X-X')}) + \mathbb{E}(e^{-t(X-X')})\right) \\
&= \mathbb{E}(\cosh(t(X-X'))) \\
&= \sum_{q \geq 0} \frac{t^{2q}\mathbb{E}[(X-X')^{2q}]}{(2q)!}
\end{aligned}$$

where the final step follows from Fubini's theorem, since the terms in the Taylor series of $x \mapsto \cosh(x)$ are non-negative. Moreover, $\mathbb{E}(e^{-tX}) \geq 1$ by Jensen's inequality. Hence by applying the inequality $(a+b)^r \leq \max(1, 2^{r-1})(a^r + b^r)$ for every $a, b, r \geq 0$ we have for every $t \in \mathbb{R}$

$$\begin{aligned}
\mathbb{E}(e^{tX}) \leq \mathbb{E}(e^{-tX})\mathbb{E}(e^{tX}) = \mathbb{E}(e^{t(X-X')}) &= \sum_{q\geq 0} \frac{t^{2q}\mathbb{E}[(X-X')^{2q}]}{(2q)!} \\
&\leq \sum_{q\geq 0} \frac{t^{2q}\mathbb{E}[(|X|+|X'|)^{2q}]}{(2q)!} \\
&\leq \sum_{q\geq 0} \frac{2^{2q-1}t^{2q}(\mathbb{E}(X^{2q}) + \mathbb{E}(X'^{2q}))}{(2q)!} \\
&\leq \sum_{q\geq 0} \frac{(2tC)^{2q}q!}{(2q)!} \\
&= \sum_{q\geq 0} \frac{(2tC)^{2q}}{\prod_{j=1}^{q}(q+j)} \\
&\leq \sum_{q\geq 0} \frac{(2tC)^{2q}}{\prod_{j=1}^{q}(2j)} \\
&= \sum_{q\geq 0} \frac{2^q(tC)^{2q}}{q!} = e^{2t^2C^2}.
\end{aligned}$$

□

Theorem (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent sub-Gaussian random variables with X_i having sub-Gaussian parameter σ_i^2 for $i \in [n]$. Then $\bar{X} := n^{-1}\sum_{i=1}^n X_i$ is sub-Gaussian with variance parameter $\frac{\bar{\sigma}^2}{n}$, where $\bar{\sigma}^2 = n^{-1}\sum_{i=1}^n \sigma_i^2$. In particular,*

$$\mathbb{P}(\bar{X} \geq x) \leq e^{-\frac{nx^2}{2\sigma^2}}$$

for every $x \geq 0$.

Proof. For every $t \in \mathbb{R}$,

$$\mathbb{E}(e^{t\bar{X}}) = \prod_{i=1}^n \mathbb{E}(e^{\frac{t}{n}X_i}) \leq \prod_{i=1}^n e^{\frac{t^2\sigma_i^2}{2n^2}} = e^{\frac{t^2\bar{\sigma}^2}{n}}.$$

The second claim then follows from part (a) of the previous proposition. □

Remark. It is often convenient to state the conclusion of Hoeffding's inequality as an upper bound on the $(1-\delta)$th qualtile of the distribution of $\bar{X}$: for every $\delta \in [0, 1]$

$$\mathbb{P}\left(\bar{X} \geq \frac{2^{1/2}\bar{\sigma}\log^{1/2}(1/\delta)}{n^{1/2}}\right) \leq \delta.$$

Remark. Since we have the same bound on $\mathbb{P}(\bar{X} \leq -x)$ we have that

$$\mathbb{P}(|X| \geq x) \leq 2e^{-\frac{nx^2}{2\sigma^2}}$$

for every $x \geq 0$. Often, Hoeffding's inequality is stated in this (weaker) way.

Lemma (Hoeffding's Lemma)**.** *Let X be a mean 0 bounded random variable taking values in $[a, b]$. Then X is sub-Gaussian with variance parameter $(b-a)^2/4$.*

Proof. Example Sheet 1. □

Corollary. *Let $X_1, \ldots, X_n$ be independent with $\mathbb{E}(X_i) = \mu_i$ and X_i taking values in $[a_i, b_i]$ for $i \in [n]$. Writing $\bar{X} := n^{-1}\sum_{i=1}^n X_i$ and $\bar{\mu} = n^{-1}\sum_{i=1}^n \mu_i$, we have*

$$\mathbb{P}(\bar{X} - \bar{\mu} \geq x) \leq \exp\left(-\frac{2n^2x^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

for every $x \geq 0$.

Proof. By Hoeffding's Lemma and Hoeffding's inequality, $\bar{X} - \bar{\mu}$ is sub-Gaussian with variance parameter $\sum_{i=1}^n \frac{(b_i-a_i)^2}{4n^2}$. □

The bound in the previous corollary may be loose when the variance of X_i is small by comparison with $(b_i - a_i)^2$. This happens for instance when $X_i \sim$ Bernoulli(p_i) with p_i close to 0 or 1. In such circumstances, Bennett's inequality (which applies when the random variables have a bounded right tail) or Bernstein's inequality (which applied under a weaker integrability condition) may be preferred.

Theorem (Bennett's inequality)**.** *Let $X_1, \dots, X_n$ be independent with $\mathbb{E}X_i = 0$ and $X_i \leq b$ for some $b > 0$ and all $i \in [n]$. Write $S = \sum_{i=1}^n X_i$ and assume that $\nu = n^{-1}\sum_{i=1}^n Var(X_i) \in (0, \infty)$, and define $\varphi : \mathbb{R} \to \mathbb{R}$ by $\varphi(u) = e^u - 1 - u$. Then, for every $t \geq 0$*

$$\log \mathbb{E}(e^{tS}) \leq \frac{n\nu}{b^2}\varphi(bt).$$

Hence, defining $h : [0, \infty) \to [0, \infty)$ by $h(u) = (1+u)\log(1+u) - u$, we have for every $x \geq 0$ that

$$\mathbb{P}\left(\frac{S}{n} \geq x\right) \leq \exp\left(-\frac{n\nu}{b^2}h\left(\frac{bx}{2}\right)\right).$$

Proof. Define $g : \mathbb{R} \to \mathbb{R}$ by

$$g(u) = \begin{cases} \frac{\varphi(u)}{u^2} & \text{if } u \neq 0 \\ \frac{1}{2} & \text{if } u = 0 \end{cases}.$$

We claim that g is increasing. To see this, define $G : \mathbb{R} \to \mathbb{R}$ by

$$G(u) = \int_0^1 e^{su}\mathrm{d}s = \begin{cases} \frac{e^u - 1}{u} & \text{if } u \neq 0 \\ 1 & \text{if } u = 0 \end{cases}.$$

Then G is convex, because if $\lambda \in (0,1)$ and $u, v \in \mathbb{R}$ we have

$$\begin{aligned} G(\lambda u + (1-\lambda)v) &= \int_0^1 e^{s(\lambda u + (1-\lambda)v)}\mathrm{d}s \\ &\leq \lambda \int_0^1 e^{su}\mathrm{d}s + (1-\lambda)\int_0^1 e^{sv}\mathrm{d}s \\ &= \lambda G(u) + (1-\lambda)G(v) \end{aligned}$$

by convexity of $u \mapsto e^{su}$. Hence

$$g(u) = \begin{cases} \frac{G(u) - G(0)}{u} & \text{if } u \neq 0 \\ G'(0) & \text{if } u = 0 \end{cases}$$

is increasing, as required.

It follows that for every $t \geq 0$,

$$e^{tX_i} - 1 - tX_i \leq X_i^2 \frac{e^{bt} - 1 - bt}{b^2} = X_i^2 \frac{\varphi(bt)}{b^2}.$$

Hence for every $t \geq 0$

$$\begin{aligned} \mathbb{E}(e^{tS}) = \sum_{i=1}^n \log \mathbb{E}(e^{tX_i}) &\leq \sum_{i=1}^n \log\left(1 + \frac{\mathbb{E}(X_i^2)\varphi(bt)}{b^2}\right) \\ &\leq n \log\left(1 + \frac{\nu}{b^2}\varphi(bt)\right) \qquad \text{(Jensen)} \\ &\leq \frac{n\nu}{b^2}\varphi(bt) \end{aligned}$$

where the final inequality uses the fact $\log(1+x) \leq x$ for $x > -1$.

Hence, by a Chernoff bound

$$\mathbb{P}\left(\frac{S}{n} \geq x\right) = \mathbb{P}(S \geq nx) \leq \inf_{t \geq 0} e^{-ntx + \frac{n\nu}{b^2}\varphi(bt)}.$$

But if $f(t) = -tx + \frac{\nu}{b^2}\varphi(bt)$ then $f'(t) = -x + \frac{\nu}{b}(e^{bt} - 1)$, so its infimum is attained at $t^* = \frac{1}{b}\log\left(1 + \frac{bx}{\nu}\right) \geq 0$, and

$$\begin{aligned}
\mathbb{P}\left(\frac{S}{n} \geq x\right) &\leq \exp\left[-\frac{nx}{b}\log\left(1+\frac{bx}{\nu}\right) + \frac{n\nu}{b^2}\left\{\left(1+\frac{bx}{\nu}\right) - 1 - \log\left(1+\frac{bx}{\nu}\right)\right\}\right] \\
&\leq \exp\left[-\frac{n\nu}{b^2}\left\{\left(1+\frac{bx}{\nu}\right)\log\left(1+\frac{bx}{\nu}\right) - \frac{bx}{\nu}\right\}\right] \\
&= \exp\left\{-\frac{n\nu}{b^2}h\left(\frac{bx}{\nu}\right)\right\}
\end{aligned}$$

for every $x \geq 0$, as required. □

Before we can state Bernstein's inequality, we first introduce the notion of sub-Gamma random variables.

Definition. A random variable X with $\mathbb{E}X = 0$ is *sub-Gamma in the right tail* with variance factor $\sigma^2 > 0$ and scale parameter $c > 0$ if

$$\log \mathbb{E}(e^{tX}) \leq \frac{\sigma^2 t^2}{2(1-ct)}$$

for all $t \in [0, 1/c)$.

Proposition.

(a) Let X be sub-Gamma in the right tail with variance parameter σ^2 and scale parameter c. Then

$$\mathbb{P}(X \geq x) \leq e^{-\frac{x^2}{2(\sigma^2 + cx)}}$$

for $x \geq 0$.

(b) Let X be a random variable with $\mathbb{E}X = 0$, $\mathrm{Var}(X) \leq \sigma^2$ and $\mathbb{E}(X_+^q) \leq \frac{q!}{2}\sigma^2 c^{q-2}$ for all integers $q \geq 3$. Then X is sub-Gamma in the right tail with variance factor σ^2 and scale parameter c.

Proof.

(a) By a Chernoff bound, for all $x \geq 0$ we have

$$\mathbb{P}(X \geq x) \leq \inf_{t \in [0, 1/c)} e^{-tx + \frac{\sigma^2 t^2}{2(1-ct)}} \leq e^{-\frac{x^2}{2(\sigma^2 + cx)}}$$

where the final inequality follows by setting $t = \frac{x}{\sigma^2 + cx} \in [0, 1/c)$.

(b) Recall from the proof of the previous theorem (Bennett's inequality) that $\varphi(u) \leq u^2/2$ for $u \leq 0$. It follows that

$$\varphi(u) = e^u - 1 - u \leq \frac{u^2}{2} + \sum_{q \geq 3} \frac{u_+^q}{q!} \text{ for all } u \in \mathbb{R}.$$

Hence, for $t \in [0, 1/c)$, by Fubini's theorem

$$\begin{aligned} \log \mathbb{E}(e^{tX}) \leq \mathbb{E}(e^{tX}) - 1 &= \mathbb{E}\varphi(tX) && \text{(since } \mathbb{E}X = 0) \\ &= \frac{t^2}{2}\mathbb{E}(X^2) + \sum_{q \geq 3} \frac{t^q \mathbb{E}(X_+^q)}{q!} \\ &\leq \frac{\sigma^2}{2} \sum_{q \geq 2} t^q c^{q-2} \\ &= \frac{\sigma^2 t^2}{2(1-ct)} \end{aligned}$$

as required.

□

Finally, we are in a position to state Bernstein's inequality.

Theorem (Bernstein's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent, mean 0 random variables with $\frac{1}{n}\sum_{i=1}^n Var(X_i) \leq \sigma^2$ and $\frac{1}{n}\sum_{i=1}^n \mathbb{E}[(X_i)_+^q] \leq \frac{q!}{2}\sigma^2 c^{q-2}$ for some $\sigma, c > 0$. Then $S = \sum_{i=1}^n X_i$ is sub-Gamma in the right tail with variance factor $n\sigma^2$ and scale parameter c. In particular, writing $\bar{X} = n^{-1}\sum_{i=1}^n X_i$, we have for every $x \geq 0$ that*

$$\mathbb{P}(\bar{X} \geq x) \leq \exp\left(-\frac{nx^2}{2(\sigma^2 + cx)}\right).$$

Proof. As in the proof of (b) of the previous proposition, for every $t \in [0, 1/c)$,

$$\begin{aligned} \log \mathbb{E}(e^{tS}) &\leq \sum_{i=1}^n \left[\frac{t^2}{2}\mathbb{E}(X_i^2) + \sum_{q \geq 3} \frac{t^q}{q!}\mathbb{E}[(X_i)_+^q]\right] \\ &\leq \frac{n\sigma^2}{2} \sum_{q \geq 2} t^q c^{q-2} && \text{(Fubini)} \\ &= \frac{n\sigma^2}{2(1-ct)} \end{aligned}$$

as required. By part (a) of the previous proposition, for every $x \geq 0$ we have

$$\mathbb{P}(\bar{X} \geq x) = \mathbb{P}(S \geq nx) \leq \exp\left(-\frac{nx^2}{2(\sigma^2 + cx)}\right)$$

as required. □

Remark. If in addition we have $X_i \leq b$ for some $b > 0$ and all $i \in [n]$, then for every integer $q \geq 3$,

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[(X_i)_+^q] \leq \frac{b^{q-2}}{n}\sum_{i=1}^{n}\mathbb{E}[X_i^2] \leq \sigma^2 b^{q-2} \leq \frac{q!}{2}\sigma^2\left(\frac{b}{3}\right)^{q-3}.$$

Thus we can take $c = b/3$. In that case, Bernstein's inequality can be deduced from Bennett's inequality, since

$$h(u) \geq \frac{u^2}{2(1+u/3)}$$

for all $u > 0$; see Example Sheet.

2 Kernel density estimation

2.1 Introduction

Let $X_1, \ldots, X_n$ be independent, real-valued random variables with density f. The oldest, and most commonly used non-parametric density estimator is the *histogram*, usually formed by dividing the real lines into equal-sized intervals, known as *bins*.

If I_x denotes the bin for which $x \in \mathbb{R}$ belongs, and $b > 0$ denotes the *binwidth*, then the histogram density estimator of f is $\hat{f}_n^H = \hat{f}_{n,b}^H$, where

$$\hat{f}_n^H(x) = \frac{1}{nb}\sum_{i=1}^{n}\mathbb{1}\{i \in I_x\}.$$

Drawbacks of histograms include:

- Difficulties of choosing the binwidth and the positioning of the bin edges;
- Suboptimal theoretic performance;
- Difficulties of graphical display in the multivariate case.

2.2 The univariate kernel density estimator

A *kernel* is a Borel-measurable function $k : \mathbb{R} \to \mathbb{R}$ satisfying $\int_{-\infty}^{\infty} k(x)\mathrm{d}x = 1$. A *univariate kernel density estimator* of f is of the form $\hat{f}_n = \hat{f}_{n,h,k}$, where

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)$$

where k is a kernel and $h > 0$ is called the *bandwidth*. It is convenient to define the *scaled kernel* $k_h(\cdot) := h^{-1}k(\frac{\cdot}{h})$, so that

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} k_h(x - X_i).$$

Typically in practice, k is chosen to be non-negative (though see later discussion), which ensures that k (and hence $\hat{f}_n$) is a density; often k is chosen to be symmetric about 0.

The main intuition is that, perhaps surprisingly, the choice of kernel is much less important than the choice of bandwidth.

2.3 Mean squared error of kernel estimators

If we think of $\hat{f}_n(x)$ as a point estimate of $f(x)$, then it is natural to try to choose h and k to minimise the *mean squared error* (MSE), defined by

$$\mathrm{MSE}[\hat{f}_n(x)] = \mathbb{E}[(\hat{f}_n(x) - f(x))^2].$$

This is often preferred to alternatives, such as the mean absolute error, due to its appealing decomposition into variance and squared bias terms:

$$\begin{aligned}\mathrm{MSE}(\hat{f}_n) &= \mathbb{E}[(\hat{f}_n(x) - \mathbb{E}[\hat{f}_n(x)])^2] + [\mathbb{E}(\hat{f}_n(x)) - f(x)]^2 \\ &= \mathrm{Var}(\hat{f}_n(x)) + \mathrm{Bias}^2\hat{\mathrm{f}}_{\mathrm{n}}(\mathrm{x}).\end{aligned}$$

The *convolution* of Borel-measurable functions $g_1, g_2 : \mathbb{R} \to \mathbb{R}$ is $g_1 * g_2$, where

$$(g_1 * g_2)(x) = \int_{\mathbb{R}} g_1(x - z)g_2(z)\mathrm{d}z,$$

whenever this integral exists. We can now compute

$$\begin{aligned}\mathrm{Bias}\hat{f}_n(x) = \mathbb{E}[k_h(x - X_1)] - f(x) &= \int_{-\infty}^{\infty} k_h(x - z)f(z)\mathrm{d}z - f(x) \\ &= (k_h * f)(x) - f(x).\end{aligned}$$

Now, for $i \in [n]$, let $\xi_i = \xi_i(x) = k_h(x - X_i)$. Then $\xi_1, \ldots, \xi_n$ are iid, so

$$\begin{aligned}\mathrm{Var}\hat{f}_n(x) = \frac{1}{n}\mathrm{Var}(\xi_1) &= \frac{1}{n}[\mathbb{E}[\xi_1^2] - \mathbb{E}^2[\xi_1]] \\ &= \frac{1}{n}\left[(k_h^2 * f)(x) - (k_h * f)^2(x)\right].\end{aligned}$$

Hence

$$\mathrm{MSE}[\hat{f}_n(x)] = \frac{1}{n}\left[(k_h^2 * f)(x) - (k_h * f)^2(x)\right] + \left[(k_h * f)(x) - f(x)\right]^2.$$

Often however, we prefer to choose h and k based on how well $\hat{f}_n$ estimates f as a function, as measured by the *mean integrated squared error* (MISE), defined by

$$\mathrm{MISE}(\hat{f}_n) = \mathbb{E}\left[\int_{\mathbb{R}} [\hat{f}_n(x) - f(x)]^2 \mathrm{d}x\right].$$

Since the integrand is non-negative, we have by Fubini's theorem that

$$\begin{aligned}\mathrm{MISE}(\hat{f}_n) &= \int_{\mathbb{R}} \mathrm{MSE}(\hat{f}_n(x))\mathrm{d}x \\ &= \frac{1}{n}\int_{\mathbb{R}} \frac{1}{n}\left[(k_h^2 * f)(x) - (k_h * f)^2(x)\right]\mathrm{d}x + \int_{\mathbb{R}} \left[(k_h * f)(x) - f(x)\right]^2 \mathrm{d}x.\end{aligned}$$

Although this expression is exact, it depends on the bandwidth in a complicated way. So we therefore seek finite-sample bounds to clarify this dependence.

2.4 Bounds on the variance and bias

For a kernel k, we write $R(k) := \int_{\mathbb{R}} k^2(u)\mathrm{d}u$.

Proposition. Let $\hat{f}_n$ denote a KDE with bandwidth h and kernel k constructed from $X_1, \ldots, X_n \sim^{\mathrm{iid}} f$. Then for every $x \in \mathbb{R}$,

$$\mathrm{Var}\hat{f}_n(x) \leq \frac{1}{nh} R(k)\|f\|_\infty.$$

Proof. We have

$$\begin{aligned}\mathrm{Var}\hat{f}_n(x) &= \frac{1}{n}\mathrm{Var}(\xi_1) \\ &\leq \frac{1}{n}\mathbb{E}(\xi_1^2) \\ &\leq \frac{1}{nh^2}\int_{\mathbb{R}} k^2\left(\frac{x-z}{h}\right) f(z)\mathrm{d}z \\ &= \frac{1}{nh}\int_{\mathbb{R}} k^2(u) f(x-uh)\mathrm{d}u \qquad (u = \tfrac{x-z}{h}) \\ &\leq \frac{1}{nh} R(k)\|f\|_\infty.\end{aligned}$$

□

To study the bias, we will need conditions on both f and k.

Definition. Let I be an interval, let $\beta, L > 0$ and let $m := \lceil \beta \rceil - 1$. The *Hölder class* $\mathcal{H}(\beta, L)$ on I is the set of m-times differentiable functions $f : I \to \mathbb{R}$ satisfying

$$|f^{(m)}(x) - f^{(m)}(x')| \leq L|x - x'|^{\beta - m}$$

for all $x, x' \in I$. Where I is unspecified we will take $I = \mathbb{R}$.