Handout 11a - Towards distribution-freeness

1 Empirical processes indexed by functions

So far, we have have seen that one can perform goodness-of-fit for univariate distributions considering the empirical process

$$v_n(x; F) = \sqrt{n}[F_n(x) - F(x)] \tag{1}$$

which can be seen as a process indexed by intervals of the type $(-\infty, x]$. Whereas, to test multivariate distributions, we rely on the (multivariate) empirical process

$$v_n(\boldsymbol{x}; F) = \sqrt{n}[F_n(\boldsymbol{x}) - F(\boldsymbol{x})], \quad \boldsymbol{x} = (x_1, \dots, x_p) \tag{2}$$

which can be seen as a process indexed by the sets $(-\infty, x_1] \times (-\infty, x_2] \times \dots (-\infty, x_p]$. All of this can be further generalized by considering the empirical processes indexed by functions. For our purposes, we will consider functions $\phi \in \mathcal{L}(F) \subset L^2(F)$. Where $L^2(F)$ is the set of square integrable functions w.r.t. F, i.e.,

$$\phi: \quad \int \phi^2_{(t)} \underbrace{dF(t)}_{f(t)\,dt} = <\phi, \phi>_F = E_F[\phi^2] < +\infty$$

whereas $\mathcal{L}(F)$ is a subset of $L^2(F)$ which collects functions which not only are square integrable w.r.t F but are also orthogonal to 1 or, in other words, have mean zero under F, i.e., Finally,

$$\int \phi(t) \cdot 1 \, dF(t) = <\phi, 1>_F = E_F[\phi] = 0$$

the empirical process indexed by $\phi \in \mathcal{L}(F)$ is defined as

$$V_n(\phi; F) = \int \phi(x)\, d V_n(x; F)$$

STOCHASTIC INTEGRAL

$$= \int \phi(x)\, d \sqrt{n}\,[F_n(x) - F(x)]$$

$$= \sqrt{n}\left[\int \phi(x)\, dF_n(x) - \int \phi(x)\, dF(x)\right] \quad \int \phi(x)\, dF(x) = 0$$

$$= \frac{1}{\sqrt{n}} \sum_{i=1}^n \phi(x_i) \qquad \int \phi(x)\, dF_n(x) = \frac{1}{n}\sum_{i=1}^n \phi(x_i)$$

Let's rewrite the multivariate empirical process in equation 2 in this form:

$$V_n(x; F) = \sqrt{n}\,[F_n(x) - F(x)]$$
$$= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\left[\mathbb{1}_{\{x_i \le \underline{x}\}} - F(\underline{x}) \right]}_{= \phi_{\underline{x}}(x_i)} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_x(x_i) = V_n(\phi_x; F)$$

- **What are its mean and covariance function?**

$$E_F\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_x(x_i)\right] \underset{\text{i.i.d}}{=} \sqrt{n}\, E_F[\phi_x(x_1)] = \sqrt{n} < \phi_x, 1 >_F = 0$$

$$E_F\left[\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_x(x_i) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \phi_{x'}(x_i)\right] = \cancel{\frac{1}{n}} \left\{ \cancel{\sum_{i=1}^{n}} E_F[\phi_x(x_i)\phi_{x'}(x_i)] + 2 \sum_{i'<i} \underbrace{E_F[\phi_x(x_i)\phi_{x'}(x_{i'})]}_{=0} \right\}$$

$$= < \phi_x, \phi_{x'} >_F$$
$$= F(x \wedge x') - F(x)F(x') < +\infty$$

$$\Rightarrow \phi_x \in \mathcal{L}(F)$$

- **What is the limiting distribution of $v_n(\phi_x; F)$?**

F-Brownian Bridge

which is a Gaussian process... and thus it is fully characterized by its mean and covariance $< \phi_x, 1 >_F$ and $< \phi_x, \phi_{x'} >_F$.

2 Constructing distribution free tests

All the consideration above hold for any arbitrary distribution function. To perform distribution-free goodness-of-fit we will consider a *reference distribution Q*, and the respective empirical process:

$$V_n(\psi_x; Q) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \psi_x(x_i) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \underbrace{\left[\mathbb{1}_{\{x_i \le x\}} - Q(x)\right]}_{\psi_x(x_i)}$$

with mean $< \psi_x, 1 >_Q = 0$

covariance $< \psi_x, \psi_{x'} >_Q < +\infty$ $\qquad \psi_x \in \mathcal{L}(Q)$

$\tilde{\phi}_x(x_i)$

But how do we use it? Instead of testing F directly by taking functionals of $v_n(\phi_x; F)$ (see Handout 11), we will construct another empirical process, namely $v_n(K l\psi_x; F)$ (to be defined in a minute) which, under F, has the same limiting distribution of $v_n(\psi_x; Q)$, under Q. Our test statistics will correspond to functionals of $v_n(K l\psi_x; F)$ and, under F, will have the same distribution as functionals of $v_n(\psi_x; Q)$, under Q. This can be done by exploiting the so-called *Khmaladze-2 (K-2) transform* and which consists of the following steps:

Step 1 - Map ψ_x into $L^2(F)$ via the isometry

$\mathcal{L}(Q)$

$$l(t) = \sqrt{\frac{q(t)}{f(t)}},$$

where q and f are the densities of Q and F, respectively. Obtain $l\psi_x \in L^2(F)$. To see that:

$$\langle l\psi_x, l\psi_x \rangle_F = \int l^2(t)\,\psi_x^2(t)\,\underbrace{dF(t)}_{f(t)dt} = \int \frac{q(t)}{\cancel{f(t)}}\,\psi_x^2(t)\,\cancel{f(t)}\,dt$$

$$= \int \psi_x^2(t)\,q(t)\,dt = \int \psi_x^2(t)\,dQ(t) = \langle \psi_x, \psi_x \rangle_Q < +\infty$$

$l\psi_x \in L^2(F)$

Do the $l\psi_x \in \mathcal{L}(F)$?

$$\langle l\psi_x, 1 \rangle_F = \int l(t)\,\psi_x(t)\,dF(t)$$

$$= \int \sqrt{\frac{q(t)}{f(t)}}\,\psi_x(t)\,f(t)\,dt$$

$$= \int \psi_x(t)\,\sqrt{f(t)\,q(t)}\,dt \neq 0$$

$$\Rightarrow l\psi_x \notin \mathcal{L}(F)$$

Step 2 - Map $l\psi_x$ $\in L^2(F)$ into $\mathcal{L}(F)$ by means of the the unitary operator

$$K = I - \frac{1-l}{1-\langle l, 1\rangle_F}\langle 1-l, \cdot\rangle_F, \tag{3}$$

it preserves the inner product

$\langle Kx, Ky\rangle = \langle x, y\rangle$

and obtain $Kl\psi_x \in \mathcal{L}(F)$. To see that:

$$K\ell\psi_x = K(\ell(t)\psi_x(t))$$

$$= \ell(t)\psi_x(t) - \frac{1-\ell(t)}{1-\int \ell(t)\,dF(t)}\int (1-\ell(t))\ell(t)\psi_x(t)\,dF(t)$$

$$= \ell(t)\psi_x(t) - \frac{1-\ell(t)}{1-\int \ell(t)\,dF(t)}\left[\int \ell(t)\psi_x(t)\,dF(t) - \underbrace{\int \ell^2(t)\psi_x(t)\,dF}_{=\langle \ell^2\psi, 1\rangle_F = \langle \psi, 1\rangle_Q = 0}\right]$$

$$= \ell(t)\psi_x(t) - \frac{1-\ell(t)}{1-\int \ell(t)\,dF(t)}\int \ell(t)\psi_x(t)\,dt$$

is it in $L^2(F)$? yes

$$\langle K\ell\psi_x, K\ell\psi_x\rangle_F = \langle \ell\psi_x, \ell\psi_x\rangle_F = \langle \psi_x, \psi_{x'}\rangle_Q < +\infty$$

is $K\ell\psi$ in $\mathcal{L}(F)$?

$$\langle K\ell\psi_x, 1\rangle_F = \int \ell\psi_x\,dF - \frac{\cancel{1-\int \ell\,dF}}{\cancel{1-\int \ell\,dF}}\int \ell\psi_x\,dF = 0$$

$\emptyset$

It follows that $v_n(Kl\psi_{\boldsymbol{x}}; F)$ and $v_n(\psi_{\boldsymbol{x}}; Q)$, converge to a Gaussian process $v(\psi_{\boldsymbol{x}}; Q)$ with mean and covariance:

MEAN $\quad \langle K\ell\psi_x, 1\rangle_F = \langle \psi_x, 1\rangle_Q = 0$

COVARIANCE $\quad \langle K\ell\psi_x, K\ell\psi_{x'}\rangle_F = \langle \ell\psi_x, \ell\psi_{x'}\rangle_F = \langle \psi_x, \psi_{x'}\rangle_Q < +\infty$

Hence we construct an entire family of test statistics for testing $H_0 : \boldsymbol{X} \sim F$ versus $H_1 : \boldsymbol{X} \not\sim F$, i.e.,

- **Kolmogorov's statistics:** $\sup |v_n(Kl\psi_{\boldsymbol{x}}; F)| \xrightarrow{d} \sup |v(\psi_{\boldsymbol{x}}; Q)|$.
- **Cramer von Mises statistics:** $\int |v_n(Kl\psi_{\boldsymbol{x}}; F)|^2 dQ(\boldsymbol{x}) \xrightarrow{d} \int |v(\psi_{\boldsymbol{x}}; Q)|^2 dQ(\boldsymbol{x})$.
- **Anderson-Darling statistics:** $\int |\frac{v_n(Kl\psi_{\boldsymbol{x}};F)}{\sqrt{Q(\boldsymbol{x})(1-Q(\boldsymbol{x}))}}|^2 dQ(\boldsymbol{x}) \xrightarrow{d} \int |\frac{v(\psi_{\boldsymbol{x}};Q)}{\sqrt{Q(\boldsymbol{x})(1-Q(\boldsymbol{x}))}}|^2 dQ(\boldsymbol{x})$.

where the convergence is intended as $n \to \infty$, under H_0. So, for sufficiently large n, a valid testing procedure consists of simulating the distribution of the functionals of $v_n(\psi_{\boldsymbol{x}}; Q)$ under Q and using it to assign significance to the values of the functionals of $v_n(Kl\psi_{\boldsymbol{x}}; F)$ observed on the data. For instance, if we decide to use Kolmogorov's statistics, we simulate the distribution of $D_Q = \sup_{\psi_{\boldsymbol{x}}} |v_n(\psi_{\boldsymbol{x}}; Q)|$ under Q. This will give us our null distribution. We then evaluate evaluate $D_{obs} = \sup_{\psi_{\boldsymbol{x}}} |v_n(Kl\psi_{\boldsymbol{x}}; F)|$ on the data observed. Our p-value is:

$$P(D_Q \geq D_{obs})$$