Lecture 8 (Continued from Lecture 7)

Goodness of Fit Testing in L_2 for the density Model

Let $X_1, X_2, \ldots, X_n$ be iid samples from a density η on $([0,1], \mathcal{B}[0,1])$. Consider the hypothesis testing problem

$$H_0: \eta \equiv 1 \quad vs \quad H_1 := \{\eta \in \mathcal{P}(\alpha, M, B) : \|\eta - 1\|_2 \geq r_n\}$$

where $\|\cdot\|_2$ stands for the L_2-norm on $[0,1]$. Problems of this kind is often called a goodness of fit problem since we are trying to understand whether the uniform distribution on $[0,1]$ fits the data better compared to set of "rougher" densities.

Let us put ourselves in the context of the general discussion above. Here $Z^{(n)} = (X_1, \ldots, X_n)$, $\Omega_n = [0,1]^n$ and $\mathcal{A}_n = \mathcal{B}[0,1]^n$. Take $\Theta = \mathcal{P}(\alpha, M, B)$ with $B > 1$, and for any $\theta \in \Theta$ identify θ with η. Then $P_\eta^{(n)}(A) = \int_A \prod_{i=1}^n \eta(x_i)\, dx_i$ for any $\eta \in \Theta$.

Also consider $\Theta \subseteq L_2[0,1]$ and thus $(\Theta, \underset{d}{\|\cdot\|_2})$ is a metric space

The testing problem is then,

$$H_0 = \{\eta_0 \equiv 1\} \quad vs \quad H_1(d, r_n) := \{\eta \in \Theta : \inf_{\eta' \in H_0} d(\eta, \eta') = \|\eta - 1\|_2 \geq r_n\}$$

Therefore to solve this problem we need to address both (i) (Upper Bound) and (ii) (Lower Bound) of our Definition of minimax optimal d-separation.

(i) Upper Bound

The crux of the argument relies on noting that
$\eta \in H_0 \Longleftrightarrow \|\eta - 1\|_2^2 = 0$ and $\eta \in H_1 \Longleftrightarrow \eta \in \mathcal{P}(\alpha, M, B)$ & $\|\eta - 1\|_2^2 \geq r_n^2$

$\therefore$ If we can estimate $\|\eta - 1\|_2^2$ for $\eta \in \mathcal{P}(\alpha, M, B)$, we can hope to tell from it whether the value is "close to 0 or not".

Let $x^*(\eta) = \|\eta - 1\|_2^2 = \|\eta\|_2^2 - 1 = x(\eta) - 1$ where
$x(\eta) = \int \eta^2(x)\,dx$

Now note that we have already considered the estimation of $x(\eta)$ with our candidate estimator

$$\hat{x}_n = \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} K_{V_{\dot{j}}}(X_{i_1}, X_{i_2})$$

$$= \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \left[\varphi_{00}(X_{i_1}) \varphi_{00}(X_{i_2}) + \sum_{l=0}^{\dot{j}} \sum_{k=0}^{2^{\dot{j}}-1} \psi_{lk}(X_{i_1}) \psi_{lk}(X_{i_2}) \right]$$

$$= \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \sum_{l=-1}^{\dot{j}} \sum_{k} \psi_{lk}(X_{i_1}) \psi_{lk}(X_{i_2}) \qquad \text{where } \psi_{-10} = \varphi_{00}$$

Therefore our estimator of $x^*(\eta)$ is

$$\hat{x}_n = \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} K_{V_{\dot{j}}}(X_{i_1}, X_{i_2}) - 1$$

$$= \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \sum_{l=-1}^{\dot{j}} \sum_{k=0}^{(2^{\dot{j}}-1) \vee 0} \left(\psi_{lk}(X_{i_1}) - \langle \psi_{lk}, \eta_0 \rangle \right) \left(\psi_{lk}(X_{i_2}) - \langle \psi_{lk}, \eta_0 \rangle \right)$$

where $\eta_0 \equiv 1$ corresponds to the uniform density

- Since $\hat{\chi}_n^*$ is a minimax optimal estimator of $\chi^*(\eta)$ over $\mathcal{P}(\alpha, M, B)$ for properly chosen j, we expect $\hat{\chi}^*$ to be "small" under the null hypothesis and "comparatively larger" under the alternative hypothesis. Therefore we will want to reject H_0 for larger values of $\hat{\chi}_n^*$

Question: How large a value of $\hat{\chi}_n^*$ should makes us reject H_0?

Idea: The answer to this question is of course guided by the necessity to control the type I error (such that under the alternative one still beats the threshold)

Fix $\alpha > 0$.

$\therefore$ Let our test be $\Psi_n \equiv I(|\hat{\chi}_n^*| \geq t_n)$ and our job is to determine the "smallest t_n" s.t $\sup_{\eta \in H_0} P_\eta(\Psi_n = 1) \to 0$ as $n \to \infty$.

By Chebychev's inequality, $P_{\eta_0}(\Psi_n = 1) = P_{\eta_0}(|\hat{\chi}_n^*| \geq t_n) \leq \dfrac{E_{\eta_0}(\hat{\chi}_n^{*2})}{t_n^2}$

So, we try to find t_n s.t. $E_{\eta_0}(\hat{\chi}_n^{*2})/t_n \leq \alpha'/2$ for $\alpha < \alpha' < 1/2$.

Note that under η_0 i.e. the null hypothesis

$$\hat{\chi}_n^* = \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \sum_{l=-1}^{j} \sum_{k=0}^{2^l - 1 \vee 0} (\Psi_{lk}(x_{i_1}) - \langle \Psi_{lk}, \eta_0 \rangle)(\Psi_{lk}(x_{i_2}) - \langle \Psi_{lk}, \eta_0 \rangle)$$

is degenerate.

Therefore by a result in Lecture 4,

$$\mathbb{E}_{\eta_0}(\hat{\chi}_n^{*2}) = \frac{2}{n(n-1)}\mathbb{E}_{\eta_0}\Big\{\sum_{l=-1}^{j}\sum_{k=0}^{2^l-1\vee 0}(\psi_{lk}(X_{i_1}) - \langle\psi_{lk},\eta_0\rangle)(\psi_{lk}(X_{i_2}) - \langle\psi_{lk},\eta_0\rangle)\Big\}^2 \quad (*)$$

To analyze the above, note that, if $e_1, e_2, \ldots$ is an o.n.b. of ~~$L_2(\eta_0)$~~ $L_2[0,1]$ then (using degeneracy)

$$\mathbb{E}_{\eta_0}\Big(\sum_{l=1}^{M}(e_l(X_1) - \langle e_l,\eta_0\rangle)(e_l(X_2) - \langle e_l,\eta_0\rangle)\Big)^2$$

$$\leq \mathbb{E}_{\eta_0}\Big(\sum_{l=1}^{M} e_l(X_1)e_l(X_2)\Big)^2 \qquad \text{(Check This)}$$

Therefore, from $(*)$ above,

$$\mathbb{E}_{\eta_0}(\tilde{\hat{\chi}}_n^{*}) = \frac{2}{n(n-1)}\mathbb{E}_{\eta_0}\Big\{\sum_{l=-1}^{j}\sum_{k=0}^{2^l-1\vee 0}(\psi_{lk}(X_1) - \langle\psi_{lk},\eta_0\rangle)(\psi_{lk}(X_2) - \langle\psi_{lk},\eta_0\rangle)\Big\}^2$$

$$\leq \frac{2}{n(n-1)}\mathbb{E}_{\eta_0}\Big[\sum_{l=-1}^{j}\sum_{k=0}^{2^l-1\vee 0}\psi_{lk}(X_1)\psi_{lk}(X_2)\Big]^2$$

$$= \cancel{\frac{2}{n(n-1)}\sum_{l=-1}^{j}\sum_{k=0}^{2^l-1}\Big\{\mathbb{E}_{\eta_0}(\psi_{lk}^2(X))\Big\}^2 + \frac{2}{n(n-1)}\sum_{l\neq l'}\sum_{k\neq k'}\Big(\mathbb{E}_{\eta_0}(\psi_{lk}(X)\psi_{l'k'}(X))\Big)^2}$$

$$\leq \frac{2}{n(n-1)}\iint\Big(\sum_{l=-1}^{j}\sum_{k=0}^{2^l-1\vee 0}\psi_{lk}(X_1)\psi_{lk}(X_2)\Big)^2\eta_0(x_1)\eta_0(x_2)\,dx_1dx_2$$

$$\leq \frac{2\|\eta_0\|_\infty^2}{n(n-1)}\iint\Big(\sum_{l=-1}^{j}\sum_{k=0}^{2^l-1\vee 0}\psi_{lk}(X_1)\psi_{lk}(X_2)\Big)^2dx_1dx_2 \qquad (\text{if } \eta_0 \in \mathcal{F}(\alpha, M, B)$$

Here $\eta_0 \equiv 1$ so can use $B = 1$)

$$= \frac{2\|\eta_0\|_\infty^2}{n(n-1)} \times 2^j = \frac{2^{j+1}}{n(n-1)}\|\eta_0\|_\infty^2$$

Therefore we have,

$$\cancel{\mathbb{E}_{\eta_0}(}\ \mathbb{P}_{\eta_0}\left(|\hat{\chi}_n^*| \geq t_n\right) \leq \mathbb{E}_{\eta_0}\left(\hat{\chi}_n^{*2}\right)/t_n^2$$

$$= \frac{2^{j+1}\,\|\eta_0\|_\infty^2}{n(n-1)\,t_n^2} \qquad (**)$$

Letting $t_n = \sqrt{\dfrac{2^{j+2}\|\eta_0\|_\infty^2}{n(n-1)\alpha'}}$ we get

$$P_{\eta_0}\left(|\hat{\chi}_n^*| \geq t_n\right) \leq \alpha'/2$$

$\therefore$ We reject $H_0 : \{\eta_0\}$ when $|\hat{\chi}_n^*| \geq \sqrt{\dfrac{2^{j+2}\|\eta_0\|_\infty^2}{n(n-1)\alpha'}}$

We now need to show that if $r_n^2 \geq C(\alpha, M, B)\, n^{-\frac{4\alpha}{4\alpha+1}}$ then,

$$\sup_{\substack{\eta \in \mathcal{P}(\alpha, M, B), \\ \|\eta - \eta_0\|_2^2 \geq r_n^2}} \mathbb{P}_\eta(\varphi_n = 0) \leq \alpha'/2$$

To do this we need to analyze $\hat{\chi}_n^*$ under each $\eta \in \mathcal{P}(\alpha, M, B)$, $\|\eta - \eta_0\|_2^2 \geq r_n^2$ separately.

Fix $\eta \in \mathcal{P}(\alpha, M, B)$ such that $\|\eta - \eta_0\|_2^2 \geq C(M,B,\alpha)\, n^{-\frac{4\alpha}{4\alpha+1}}$

$$\mathbb{P}_\eta(\Psi_n = 0) = \mathbb{P}_\eta\left(|\hat{X}_n^*| \leq t_n\right)$$

$$= \mathbb{P}_\eta\left(|\hat{X}_n^* - \mathbb{E}_\eta \hat{X}_n^* + \mathbb{E}_\eta \hat{X}_n^*| \leq t_n\right)$$

$$\leq \cancel{\geq}\ \mathbb{P}_\eta\left(|\hat{X}_n^* - \mathbb{E}_\eta \hat{X}_n^*| \geq -t_n + |\mathbb{E}_\eta \hat{X}_n^*|\right) \qquad (*)$$

Now, $t_n = \sqrt{\dfrac{2^{j+2}\|\eta_0\|_\infty^2}{n(n-1)\alpha'}}$

$$|\mathbb{E}_\eta \hat{X}_n^*| = \mathbb{E}_\eta\left(\frac{1}{n(n-1)} \sum_{i_1 \neq i_2} K_{V_j}(X_{i_1}, X_{i_2}) - 1\right)$$

$$= \mathbb{E}_\eta\left(\frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \sum_{\ell=-1}^{j} \sum_{k=0}^{2^\ell - 1 \vee 0} \left(\psi_{\ell k}(X_{i_1}) - \langle \psi_{\ell k}, \eta_0\rangle\right)\left(\psi_{\ell k}(X_{i_2}) - \langle \psi_{\ell k}, \eta_0\rangle\right)\right)$$

$$= \sum_{\ell=-1}^{j} \sum_{k=0}^{2^\ell - 1 \vee 0} \left\{\mathbb{E}_\eta\left(\psi_{\ell k}(X) - \langle \psi_{\ell k}, \eta_0\rangle\right)\right\}^2$$

$$= \sum_{\ell=-1}^{j} \sum_{k=0}^{2^\ell - 1 \vee 0} \left\{\langle \psi_{\ell k}, \eta - \eta_0\rangle\right\}^2 = \|\Pi_{V_j}(\eta - \eta_0)\|_2^2$$

$$= \|\eta - \eta_0\|_2^2 - \|\Pi_{V_j^\perp}(\eta - \eta_0)\|_2^2$$

$$\geq C(M,B,\alpha)\, n^{-\frac{4\alpha}{4\alpha+1}} - (2M)\, 2^{-2j\alpha} \quad \text{as } \eta - \eta_0 \in \mathcal{P}(\alpha, 2M, 2B)$$

$\Rightarrow$ If $2^j \approx n^{\frac{2}{4\alpha+1}}$ i.e. $j \sim \frac{2}{4\alpha+1}\log_2 n$ we have that for sufficiently large $C(M,B,\alpha)$

$\mathbb{E}_\eta|\hat{X}_n^*| \geq t_n$. Indeed by taking $C(M,B,\alpha)$ large enough, $\mathbb{E}_\eta|\hat{X}_n^*| \geq \dfrac{C(M,B,\alpha)}{3} n^{-\frac{4\alpha}{4\alpha+1}} = \dfrac{C(M,B,\alpha)}{3} 2^{-2j\alpha}$

$$= \frac{C(M,B,\alpha)}{3}\sqrt{\frac{2^j}{n^2}}$$

Note that the crux of the above choice of j is driven by the fact $2^{-2j\alpha} = \sqrt{\frac{2^j}{n^2}}$ is solved by $2^j \sim n^{\frac{2}{4\alpha+1}}$. This is exactly similar to the bias variance tradeoff while estimating a quadratic functional. The difference is that the $\sqrt{1/n}$ term does not arise in the std. deviation due to first order degeneracy of $\hat{\chi}_n^*$.

Collecting the above calculations in $(*)$ implies,

$$\mathbb{P}_\eta(\psi_n = 0) \leq \mathbb{P}_\eta\left(|\hat{\chi}_n^* - \mathbb{E}_\eta \hat{\chi}_n^*| \geq |\mathbb{E}_\eta \hat{\chi}_n^*| - t_n\right)$$

$$= \mathbb{P}_\eta\Big(|\hat{\chi}_n^* - \mathbb{E}_\eta \hat{\chi}_n^*| \geq \underbrace{\|\Pi_{V_j}(\eta - \eta_0)\|_2^2}_{\geq \frac{C(M,B,\alpha)}{3} n^{-\frac{4\alpha}{4\alpha+1}}} - t_n\Big)$$

Now, in order to apply Chebychev's inequality, we need to understand the variance of $\hat{\chi}_n^*$, which can be done by a Hoeffding's Decomposition of $\hat{\chi}_n^* - \mathbb{E}_\eta \hat{\chi}_n^*$ under $\mathbb{P}_\eta$.

Cheek: (See chapter 11 & 12, And van Der Vaart: Asymptotic Statistics for reference, on Hoeffding's Decomposition)

$$\hat{\chi}_n^* - \mathbb{E}_\eta \hat{\chi}_n^* = \frac{2}{n} \sum_{i=1}^{n} \sum_{l=1}^{j} \sum_{k} \left(\psi_{lk}(X_i) - \langle \psi_{lk}, \eta \rangle\right) \langle \psi_{lk}, \eta - \eta_0 \rangle$$

$$+ \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \sum_{l=-1}^{j} \sum_{k=0}^{2^j - 1 \vee 0} \left(\psi_{lk}(X_i) - \langle \psi_{lk}, \eta \rangle\right)\left(\psi_{lk}(X_j) - \langle \psi_{lk}, \eta \rangle\right)$$

$$= T_1(\eta) + T_2(\eta)$$

$$\Rightarrow \mathbb{P}(\psi_n = 0) \le \mathbb{P}_\eta\left(|T_1(\eta) + T_2(\eta)| \geq \|\Pi_{V_j^*}(\eta - \eta_0)\|_2^2 - t_n\right)$$

$$\le \mathbb{P}_\eta\left(|T_1(\eta)| \geq \|\Pi_{V_j^*}(\eta-\eta_0)\|_2^2/4\right)$$

$$+ \mathbb{P}_\eta\left(|T_2(\eta)| \geq \frac{3}{4}\|\Pi_{V_j^*}(\eta-\eta_0)\|_2^2 - t_n\right)$$

$$= I_1 + I_2$$

Controlling I_1: $\mathbb{P}_\eta\left(|T_1(\eta)| \geq \|\Pi_{V_j}(\eta-\eta_0)\|_2^2/4\right)$

$$\le \frac{\mathrm{Var}_\eta T_1^*(\eta)}{\|\Pi_{V_j^*}(\eta-\eta_0)\|_2^4/16}$$

$$\mathrm{Var}_\eta(T_1^*(\eta)) = \mathrm{Var}_\eta\left(\frac{2}{n}\sum_{i=1}^n \sum_{l,k}\left(\psi_{lk}(x_i) - \mathbb{E}_\eta(\psi_{lk})\right)\right)\langle \psi_{lk}, \eta-\eta_0\rangle$$

$$= \mathrm{Var}_\eta\left(\frac{2}{n}\sum_{i=1}^n\left\{\left(\sum_{l,k}\psi_{lk}(x_i)\langle\psi_{lk},\eta-\eta_0\rangle\right) - \left(\mathbb{E}_\eta\sum_{l,k}\psi_{lk}\langle\psi_{lk},\eta-\eta_0\rangle\right)\right\}\right)$$

$$= \frac{4}{n}\mathrm{Var}_\eta\left(\sum_{l,k}\psi_{lk}(x_i)\langle\psi_{lk},\eta-\eta_0\rangle - \mathbb{E}_\eta\sum_{l,k}\psi_{lk}(x_i)\langle\psi_{lk},\eta-\eta_0\rangle\right)$$

$$\le \frac{4}{n}\mathbb{E}_\eta\left(\sum_{l,k}\psi_{lk}(x_i)\langle\psi_{lk},\eta-\eta_0\rangle\right)^2$$

$$= \frac{4}{n}\int\left(\sum_{l,k}\psi_{lk}(x)\langle\psi_{lk},\eta-\eta_0\rangle\right)^2\eta(x)\,dx$$

$$\leq \frac{4\|\eta\|_\infty^2}{n}\int\Big(\sum_{l,k}\psi_{lk}(x)\langle\psi_{lk},\eta-\eta_0\rangle\Big)^2 dx$$

$$= \frac{4\|\eta\|_\infty^2}{n}\sum_{l,k}\langle\psi_{lk},\eta-\eta_0\rangle^2 = \frac{4\|\eta\|_\infty^2}{n}\|\Pi_{V_j}(\eta-\eta_0)\|_2^2$$

$$\Rightarrow \mathbb{P}_\eta\big(|T_1(\eta)| \geq \|\Pi_{V_j}(\eta-\eta_0)\|_2^2/4\big) \leq 216\|\eta\|_\infty^2/n\|\Pi_{V_j}(\eta-\eta_0)\|_2^2$$

$$\leq \frac{216\|\eta\|_2^2}{n^{\frac{1}{4\alpha+1}}}$$

Controlling I_2: First note that

$$\frac{3}{4}\|\Pi_{V_j}(\eta-\eta_0)\|_2^2 - t_n$$

$$\geq \frac{3}{4}\big(\|\eta-\eta_0\|_2^2 - (2M)2^{-2j\alpha}\big) - \sqrt{\frac{2^{j+2}\|\eta_0\|_\infty^2}{n(n-1)\alpha}}$$

$$\geq \frac{3}{4}\big(C(M,B,\alpha)\,n^{-\frac{4\alpha}{4\alpha+1}} - (2M)2^{-2j\alpha}\big) - \sqrt{\frac{2^{j+2}B^2}{n(n-1)\alpha}}$$

$$\geq \sqrt{\frac{2^{j+3}B^2}{n(n-1)\alpha}}$$ if $2^j \sim n^{\frac{2}{4\alpha+1}}$ and $C(M,B,\alpha)$ is chosen large enough.

Therefore

$$\mathbb{P}_\eta\big(|T_2(\eta)| \geq \frac{3}{4}\|\Pi_{V_j}(\eta-\eta_0)\|_2^2 - t_n\big)$$

$$\leq \mathbb{P}_\eta\Big(\Big|\frac{1}{n(n-1)}\sum_{i_1\neq i_2}\sum_{l,k}(\psi_{lk}(X_{i_1})-\langle\psi_{lk},\eta\rangle)(\psi_{lk}(X_{i_2})-\langle\psi_{lk},\eta\rangle)\Big| \geq \sqrt{\frac{2^{j+3}B^2}{n(n-1)\alpha}}\Big)$$

which by calculations similar to (**) above on page (10) is $\leq \alpha/4$

Therefore we have proved that ~~for~~, for $C(M, B, \alpha')$ large enough, if $r_n^2 = C(M, B, \alpha') \, n^{-\frac{4\alpha}{4\alpha+1}}$, then

$$\sup_{\eta \in H_1(d, r_n)} \mathbb{P}_\eta(\psi_n = 0) \le \alpha'/2$$

This combined with our bound on type I error proves the upper bound of the theorem.

(ii) Lower Bound

The proof of the lower bound is exactly the same as the proof of lower bound for estimation of quadratic functional.

Following our stratigy, take $\Theta_0 \in H_0$ as $\Theta_0 = \eta_0$ and $\Theta_M \subseteq H_1(d, \rho_n)$ with $\rho_n << r_n = n^{-4\alpha/4\alpha+1}$ as

$$\eta_\lambda(x) = \eta_0(x) + c\, 2^{-j\alpha} \frac{2^{-j}}{a_n} \sum_k \lambda_k \psi_{jk}(x) \qquad \lambda \in \{-1, 1\}^{2^j}, \; M = 2^{2^j}$$

~~where~~ $2^j \sim n^{2/4\alpha+1}$ and $a_n \to \infty$ is s.t. $\frac{2^{-2j\alpha}}{a_n} = n^{-4\alpha/4\alpha+1}$ (this can be done since $\rho_n << n^{-4\alpha/4\alpha+1}$ and therefore these η_λ's $\in H_1(d, \rho_n)$)

The rest is an exactly similar second moment argument.

(Do it by yourself to be sure).