# Lec 11: Testing Hypotheses about Densities

Weiping Zhang

November 17, 2020

# Testing Hypotheses about Densities

- Suppose $f$ and $g$ are two possible densities for the random variable or vector $X$. We may like to test several types of hypotheses regarding these densities, each of which will be formulated as testing for

$$H_0 : f(x) = g(x) \leftrightarrow H_1 : f(x) \neq g(x)$$

- Pagan and Ullah (1999) consider several examples which we reformulate below.
  - It is sometimes desirable to test whether a nonparametrically estimated density has a particular form, say normal density.
  - Testing for symmetry of a density around some point
  - Conditional symmetry of a conditional density may be of great interest also.

- Pagan and Ullah (1999) consider several examples which we reformulate below.
  - Testing for various variants of independence such as serial independence, spatial independence, or conditional independence
  - Compare densities $f$ and $g$ that come from two different groups
- The above testing problems can be tackled by considering a widely accepted measure of global distance (closeness) between two densities $f$ and $g$. In practice, people frequently use the weighted integrated squared error:

$$I(f,g) = \int [f(x) - g(x)]^2 w(x) dx$$

where $w(x)$ is a nonnegative wight function.

- For example, if $w(x) = f(x)$ or $g(x)$, then the above error can be estimated by its sample analogue

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n}[\hat{f}(X_i) - \hat{g}(X_i)]^2$$

- Another measure of distance (affinity) between two densities is the well known Kullback-Leibler (KL) distance (information) measure introduced earlier on. Under the null hypothesis, the KL distance between $f$ and $g$ is zero and it is nonzero otherwise.

$$d(f,g) = E_f log\frac{f}{g}$$

# Comparison with a Parametric Density Function

- Consider the problem of testing

$$H_0 : f(x) = g(x;\theta)$$

  where $g(x;\theta)$ is a fully specified(known) density up to the finite dimensional parameters.

- Given data $\{X_1, \ldots, X_n\}$, let $\hat{f}(x)$ be the nonparametric kernel density estimate of $f$ and $\hat{\theta}$ be the maximum likelihood estimator for $\theta$ based upon the parametric density $g(x;\theta)$.

- Noting that

$$\begin{aligned} I(f,g) &= \int [f(x) - g(x;\theta)]^2 dx \\ &= \int f^2(x)dx + \int g^2(x;\theta)dx - 2\int f(x)g(x;\theta)dx \\ &= Ef(X) + \int g^2(x;\theta)dx - 2Eg(X;\theta) \end{aligned}$$

- Following Fan (1994), we can propose a feasible test statistic by replacing

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \hat{f}_{-i}(X_i) + \int g^2(x;\hat{\theta})dx - \frac{2}{n}\sum_{i=1}^{n} g(X_i;\hat{\theta})$$

- We can follow the proof of Theorem 4.1 of Fan (1994) to prove the following Theorem.

## Theorem

*Under some regularity conditions and $H_0$, we have*

$$T = \frac{n(h_1 \cdots h_d)^{1/2}\hat{I}}{\hat{\sigma}} \rightsquigarrow N(0,1)$$

*where* $\hat{\sigma}^2 = (n^2 h_1 \cdots h_d)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathcal{K}^2(\frac{X_i - X_j}{h})$ *and* $\mathcal{K}(\frac{X_i - X_j}{h}) = \prod_{s=1}^{d} K(\frac{X_{is} - X_{js}}{h_s})$.

# Testing for Symmetry

- To test whether a density function $f$ is symmetric around zero, we write the null and alternative hypotheses as

$$H_0 : f(x) = f(-x) \leftrightarrow H_1 : f(x) \neq f(-x)$$

- Noting that

$$\begin{aligned}
I(f,g) &= \frac{1}{2}\int [f(x) - f(-x)]^2 dx \\
&= \frac{1}{2}\int [f(x) - f(-x)] f(x) dx - \frac{1}{2}\int [f(x) - f(-x)] f(-x) dx \\
&= \int [f(x) - f(-x)] f(x) dx = \int [f(x) - f(-x)] dF(x)
\end{aligned}$$

- Ahmad and Li (1997) propose a test based upon the last functional.

- Clearly, we can estimate $I$ by

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n}[\hat{f}(X_i) - \hat{f}(-X_i)]$$

$$= \frac{1}{n^2 h_1 \cdots h_d}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\mathcal{K}(\frac{X_i - X_j}{h}) - \mathcal{K}(\frac{X_i + X_j}{h})\right]$$

- Under the null hypothesis and the standard assumption that $h_s \to 0$ and $nh_1 \cdots h_d \to \infty$, Ahmad and Li (1997) prove the following theorem

## Theorem

*Under some regularity conditions and $H_0$ we have*

$$T = \frac{n(h_1 \cdots h_d)^{1/2}(\hat{I} - c(n))}{\hat{\sigma}} \rightsquigarrow N(0,1)$$

where $\hat{\sigma}^2 = \frac{1}{4n}\sum_{i=1}^{n} \hat{f}(X_i)\|\mathcal{K}\|^2$ and $c(n) = \mathcal{K}(0)/(nh_1 \cdots h_d)$ is used to correct for finite sample bias.

# Comparison with Unknown Densities

- Comparison of two densities is important in some empirical work. For example, we may be interested in comparing income distributions across two groups, regions, or time periods.
- Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ be two samples from $d$-dimensional random vectors. Assume that $X$ and $Y$ have density $f$ and $g$ and distribution functions $F$ and $G$, respectively.
- The null hypothesis of interest is

$$H_0 : f(x) = g(x)$$

- Noticing that

$$
\begin{aligned}
I &= \int [f(x) - g(x)]^2 dx \\
&= \int f(x) dF(x) + \int g(x) dG(x) - 2 \int f(x) g(x) dx
\end{aligned}
$$

- we can propose a feasible test statistic by replacing $f, g, F$ and $G$ by $\hat{f}, \hat{g}, \hat{F}$ and $\hat{G}$, respectively, where $\hat{f} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{K}_h(X_i - x)$ and $\hat{g}(x) = \frac{1}{m}\sum_{i=1}^{m} \mathcal{K}_h(Y_i - y)$, and $\hat{F}$ and $\hat{G}$ are the empirical distributions of $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, respectively. This leads to

$$
\begin{aligned}
\hat{I} &= \int \hat{f}(x) d\hat{F}(x) + \int \hat{g}(x) d\hat{G}(x) - 2 \int \hat{f}(x) d\hat{G}(x) \\
&= \frac{1}{n} \sum_{i=1}^{n} \hat{f}(X_i) + \frac{1}{m} \sum_{i=1}^{m} \hat{g}(Y_i) - \frac{2}{m} \sum_{i=1}^{m} \hat{f}(Y_i)
\end{aligned}
$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_{h,ij}^{x} + \frac{1}{m^2}\sum_{i=1}^{m}\sum_{j=1}^{m} K_{h,ij}^{y} - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} K_{h,ij}^{xy}$$

where
$K_{h,ij}^{x} = \prod_{s=1}^{d} h_s^{-1} K((X_{is} - X_{js})/h_s)$,
$K_{h,ij}^{y} = \prod_{s=1}^{d} h_s^{-1} K((Y_{is} - Y_{js})/h_s)$,
$K_{h,ij}^{xy} = \prod_{s=1}^{d} h_s^{-1} K((X_{is} - Y_{js})/h_s)$.
The following theorem states the main result.

### Theorem

*Under some regularity conditions and $H_0$ we have*

$$T = \frac{(nmh_1 \cdots h_d)^{1/2}(\hat{I} - c(n))}{\hat{\sigma}} \rightsquigarrow N(0,1)$$

where

$$\hat{\sigma}^2 = 2nmh_1 \cdots h_d \Big\{ \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(K_{h,ij}^{x})^2}{n^4} + \sum_{i=1}^{m} \sum_{j=1}^{m} \frac{(K_{h,ij}^{y})^2}{m^4} + \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(K_{h,ij}^{xy})^2}{(nm)^2} \Big\}$$

and $c(n) = \frac{\kappa_{02}^d}{h_1 \cdots h_d}(\frac{1}{n} + \frac{1}{m})$.

- For a proof of the above result, see Li and Racine (2006). See a variant of the above test, see Li (1996).

# Testing for Independence

- Let $(X, Y)'$ be a $(p + q) \times 1$ random vector with joint cdf $F(x, y)$ and pdf $f(x, y)$. Further let $F_1(x)$ and $F_2(y)$ denote the marginal cdf of $X$ and $Y$ with marginal pdf $f_1(x)$ and $f_2(y)$, respectively. The null hypothesis of interest is

$$H_0 : f(x, y) = f_1(x) f_2(y)$$

- Observing that

$$\begin{aligned} I &= \int [f(x, y) - f_1(x) f_2(y)]^2 dx dy \\ &= \int f(x, y) dF(x, y) + \int f_1(x) dF_1(x) \int f_2(y) dF_2(y) \\ &\quad - 2 \int f_1(x) f_2(y) dF(x, y) \\ &= Ef(X, Y) + E[f_1(X)]E[f_2(Y)] - 2E[f_1(X) f_2(Y)] \end{aligned}$$

we can propose a feasible test statistic by replacing $f(X_i, Y_i)$, $f_1(X_i)$ and $f_2(Y_i)$ by their leave-one-out kernel estimators $\hat{f}_{-i}(X_i, Y_i)$, $\hat{f}_{1,-i}(X_i)$ and $\hat{f}_{2,-i}(Y_i)$. This will lead to the following expression

$$\begin{aligned}\hat{I} = \frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i, Y_i) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{f}_{-i}(X_i)\hat{f}_{2,-i}(Y_i) \\ -\frac{2}{n}\sum_{i=1}^{n}\hat{f}_{1,-i}(X_i)\hat{f}_{2,-i}(Y_i)\end{aligned}$$

where $\hat{f}_{-i}(X_i, Y_i) = \frac{1}{n-1}\sum_{j\neq i} K_{h_x}(X_j - X_i)K_{h_y}(Y_j - Y_i)$, $\hat{f}_{1,-i}(X_i) = \frac{1}{n-1}\sum_{j\neq i} K_{h_x}(X_j - X_i)$, and $\hat{f}_{2,-i}(Y_i) = \frac{1}{n-1}\sum_{j\neq i} K_{h_y}(Y_j - Y_i)$, with $K_{h_x}(X_j - X_i) = \prod_{s=1}^{p} h_{xs}^{-1} K((X_{js} - X_{is})/h_{xs})$ and $K_{h_y}(Y_j - Y_i) = \prod_{s=1}^{q} h_{ys}^{-1} K((Y_{js} - Y_{is})/h_{ys})$

Under certain conditions, Ahmad and Li (1997) prove the following theorem.

## Theorem

*Under some regularity conditions and $H_0$ we have*

$$T = \frac{n(h_{x,1} \cdots h_{x,p} h_{y,1} \cdots h_{y,q})^{1/2} \hat{I}}{\hat{\sigma}} \rightsquigarrow N(0,1)$$

*where* $\hat{\sigma}^2 = 2(n^2 h_{x,1} \cdots h_{x,p} h_{y,1} \cdots h_{y,q})^{-1} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \mathcal{K}^2(\frac{X_i - X_j}{h_x}) \mathcal{K}^2(\frac{Y_i - Y_j}{h_y})$*,*
*with, e.g.,* $\mathcal{K}(\frac{X_i - X_j}{h_x}) = \prod_{s=1}^{p} K(\frac{X_{is} - X_{js}}{h_{x,s}})$*.*

# Test for Structural Change in Densities

- The problem of testing for a structural change has generated much interest in both statistics and econometrics.
- Early study mainly focused on the case of parameter change in the parametric framework. Recently, much attention has been paid to the whole distribution or density level when testing for structural change.
- Let $\{X_t, t \geq 1\}$ be a stationary strong mixing process satisfying

$$\alpha(\tau) = \sup\{|P(A \cap B) - P(A)P(B)| : A \in \mathcal{F}_1^t, B \in \mathcal{F}_{t+\tau}^{\infty}\} \to 0$$

where $\mathcal{F}_a^b = \sigma(X_a, \ldots, X_b)$ is the $\sigma-$field generated by $X_a, \ldots, X_b$, and $1 \leq a \leq b \leq \infty$.

- We wish to test for the change of the marginal density $f$ of $\{X_t\}_{t=1}^n$. So the null hypothesis is

$$H_0 : X_1, \ldots, X_n \text{have a common marginal density} f$$

and the alternative hypothesis is

$$H_1 : \text{for some}\, s \in (0,1), X_1, \ldots, X_{\lceil ns \rceil} \text{ have a common density} f_1,$$

$$\text{and} X_{\lceil ns \rceil + 1}, \ldots, X_n \text{have a common density} f_2$$

where $\lceil a \rceil$ denotes the largest integer less than or equal to $a$, $f, f_1$ and $f_2$ are all assumed unknown.

- To test $H_0$, define

$$f_{\lceil ns \rceil}(x) = \frac{1}{\lceil ns \rceil h} \sum_{i=1}^{\lceil ns \rceil} K(\frac{x - X_i}{h})$$

and

$$f^*_{n-\lceil ns \rceil}(x) = \frac{1}{(n - \lceil ns \rceil)h} \sum_{i=\lceil ns \rceil + 1}^{n} K(\frac{x - X_i}{h})$$

- Define

$$d_n(s, x) = \Big(\frac{nh}{f_n(x)\kappa_{02}}\Big)^{1/2} \frac{\lceil ns \rceil}{n} \frac{n - \lceil ns \rceil}{n} \Big(f_{\lceil ns \rceil}(x) - f^*_{n-\lceil ns \rceil}(x)\Big)$$

provided $f_n(x) \neq 0$. If $f_n(x) = 0$, the above is defined to be zero.

Under the null $H_0$, we can define a partial sum process:

$$g_n(s,x) = \Big(\frac{f_n(x)\kappa_{02}}{nh}\Big)^{-1/2} \frac{\lceil ns \rceil}{n} \Big(f_{\lceil ns \rceil}(x) - Ef_{\lceil ns \rceil}(x)\Big)$$

$$= \Big(\frac{f_n(x)\kappa_{02}}{nh}\Big)^{-1/2} \sum_{i=1}^{\lceil ns \rceil} \Big[K(\frac{x - X_i}{h}) - EK(\frac{x - X_i}{h})\Big]$$

Then we can write

$$d_n(s,x) = g_n(s,x) - \frac{\lceil ns \rceil}{n} g_n(1,x)$$

- Lee and Na (2004) shows for fixed $x$, $\{g_n(s,x) : 0 \leq s \leq 1\}$ converge weakly to a standard Brownian motion process, which implies that $\{d_n(s,x) : 0 \leq s \leq 1\}$ converge to a Brownian bridge.

Let $x_1, \ldots, x_N$ be distant real numbers. Define

$$T_n = \max_{1 \le i \le N} \sup_{0 \le s \le 1} |d_n(s, x_i)|$$

Lee and Na (2004) prove the following theorem.

### Theorem

*Suppose the regularity conditions given in Lee and Na (2004) hold.*
*(1) Under $H_0$, as $n \to \infty, T_n \rightsquigarrow \max_{1 \le i \le N} \sup_{0 \le s \le 1} |W_i^0(s)|$, where $W_1^0, \ldots, W_N^0$ are independent Brownian bridges.*
*(2) Under $H_1$, as $n \to \infty, T_n \to \infty$ in probability, if $f_1(x_i) \neq f_2(x_i)$ for some $x_i \in \{x_1, \ldots, x_N\}$.*

Thus we reject the null if $T_n$ is large enough. In practice, one can tabulate the critical values based on simulations on Brownian bridges.

# Bootstrap test for densities

- The goodness about the above tests associated with kernel density estimates is that they are all asymptotically normally distributed.
- One should keep in mind that the asymptotic normal approximation to the exact (finite sample) distribution of the nonparametric test statistic may be poor in finite samples. Unfortunately, this is true in practice and we need to use bootstrap or some other resampling techniques to approximate the finite sample distribution of the test statistic.
- Let $T_n = T_n(X_1, \ldots, X_n)$ be a statistic of interest with cdf $G_n(x, F) = P_F(T_n \leq x)$ where $F$ is the cdf of $X_i$. Even though we know that $T_n$ is asymptotically $N(0, 1)$, we usually don't know its finite sample exact distribution.

- In this case, we can resort to a bootstrap procedure to improve the finite sample performance of the test based upon $T_n$.
- It is worth mentioning that there is no bootstrap test procedure that is universal and works for all tests.
- This is true because we have to impose the null hypothesis when we do the bootstrap test. Different null hypotheses may deserve different bootstrap testing procedure. When the data are dependent, we may also need to consider the dependence structure in the data in order to conduct a valid bootstrap test.

# Example

**Comparison with a parametric density function**

When we test for $H_0 : f(x) = g(x;\theta)$, we can impose the null hypothesis by drawing bootstrap resamples from $g(x;\hat{\theta})$, where $\hat{\theta}$ is the MLE for $\theta$. So the bootstrap testing procedure goes as follows.

- Step 1. Draw a bootstrap resample $X_1^*, \ldots, X_n^*$ from $g(x;\hat{\theta})$.
- Step 2. Use the bootstrap resample calculating the test statistic

  $$T^* = \frac{n(h_1 \cdots h_d)^{1/2}\hat{I}^*}{\hat{\sigma}^*}$$

  where $\hat{I}^*$ and $\hat{\sigma}^*$ are the same as $\hat{I}$ and $\hat{\sigma}$ except that we replace the original sample $X_1, \ldots, X_n$ by the bootstrap resample $X_1^*, \ldots, X_n^*$

- Step 3. Repeat Steps 1-2 for $B$ times and get $T_j^*, j = 1, \ldots, B$. Then we can calculate the bootstrap $p$-values as

  $$p^* = \frac{1}{B} \sum_{j=1}^{B} I(T \leq T_j^*)$$

  and reject the null hypothesis if $p^*$ is smaller than the given significance level $\alpha$.

# Example

**Testing for symmetry**

To construct a bootstrap test for $H_0 : f(x) = f(-x)$ versus $H_1 : f(x) \neq f(-x)$, we can impose the null hypothesis by bootstrapping resamples from $\{X_i, -X_i\}_{i=1}^n$. So Step 1 in the above procedures will be replaced by:

Step 1a: Draw a bootstrap resample $X_1^*, \dots, X_n^*$ from $\{X_i, -X_i\}_{i=1}^n$.

# Example

**Comparison with unknown densities**

To construct a bootstrap test for $H_0 : f(x) = g(y)$, we can impose the null as in Step 1b:

Step 1b: Draw a bootstrap resample $\{\{X_i^*\}_{i=1}^n, \{Y_i^*\}_{i=1}^m\}$ from $\{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$.

# Example

**Testing for independence**

To construct a bootstrap test for $H_0 : f(x,y) = f_1(x)f_2(y)$, we can impose the null as in Step 1c:

Step 1c: Draw a bootstrap resample $\{X_i^*\}_{i=1}^n$ from $\{X_1, \ldots, X_n\}$ and $\{Y_i^*\}_{i=1}^n\}$ independently from $\{Y_1, \ldots, Y_n\}$.