Statistics 583, Final Exam Solutions

Wellner; 6/10/2015

1. (36 points) **Define any three of** the following terms.
   (a) A Frechet - differentiable functional $T : \mathcal{F} \to \mathbb{R}$ (with respect to some metric between distribution functions) and the corresponding influence function.
   (b) The Lévy metric d_L between distribution functions. Is it *compatible* to the empirical distribution function?
   (c) A linear smoother $\hat{r}_n$ of a regression function r.
   (d) The locally-linear estimator $\hat{r}_n$ of a regression function r on $[a, b] \subset \mathbb{R}$.
   (e) A "rule-of-thumb" band-width h_n based on a normality assumption and the assumption of a twice differentiable density function f.
   (f) The Nadaraya-Watson estimator of a regression function r.

   Solution: See course notes and textbooks.

2. (36 points). **State three** of the following results or theorems:
   (a) A limit theorem for the the bootstrap empirical process $\sqrt{m}(\mathbb{F}_m^* - \mathbb{F}_n)$ when $m \wedge n \to \infty$.
   (b) A limit theorem for the general bootstrap empirical process $\mathbb{G}_n^* = \sqrt{n}(\mathbb{P}_n^* - \mathbb{P}_n)$ indexed by a class of functions $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ (with conditions specified in terms of the behavior of the empirical process $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$).
   (c) Hoeffding's exponential inequality for a sum of independent bounded random variables, with $a_i \le X_i \le b_i$ for $i = 1, \ldots, n$
   (d) Assouad's (Lower Bound) Lemma.
   (e) A limit theorem for a general "resampling without replacement" bootstrap for $T(F)$.
   (f) A theorem concerning the nonparametric bootstrap of a differentiable functional.

   Solution: See course notes and textbooks.

3. Suppose that $X_1, \ldots, X_n$ are i.i.d. with density f on $\mathbb{R}$; you may assume that f is uniformly continuous and bounded.
(a) If k is a fixed probability density on $\mathbb{R}$ which is symmetric about 0 and satisfies $\int z^2 k(z)dz = 1$, $\int k^2(z)dz < \infty$, and $h > 0$, define the kernel density estimator $\hat{f}_n$ based on k and h at a given $x \in \mathbb{R}$.
(b) Calculate $E_f(\hat{f}_n(x))$ and use it to give an expression for $\text{bias}_n(\hat{f}_n(x))$. Use your expression to explain how the bias changes as h increases.
(c) Calculate $Var_f(\hat{f}_n(x))$. If $h = h_n \to 0$ and $nh_n \to \infty$, show that

$$nh_n Var_f(\hat{f}_n(x)) \to f(x) \int k^2(z)dz \quad \text{as} \quad n \to \infty.$$

(d) Assuming that f'' exists and is continuous at x, how would you choose $h = h_n$ in order to minimize $E_f(\hat{f}_n(x) - f(x))^2$? How would you choose $h = h_n$ if you wanted your estimator to satisfy $\sqrt{nh_n}(\hat{f}_n(x) - f(x)) \to_d N(0, f(x)\int k^2(z)dz)$?

Solution: (a) Let $\mathbb{F}_n$ denote the empirical d.f. of the X_i's. Then

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

(b) Since $\hat{f}_n$ is the sum of n identically distributed terms,

$$\begin{aligned} E_f\hat{f}_n(x) &= \frac{1}{h}E_f k\left(\frac{x-X_1}{h}\right) = \frac{1}{h}\int k\left(\frac{x-y}{h}\right) f(y)dy \\ &= \int k(z)f(x - hz)dz \quad \text{by the change of variables } z = (x-y)/h. \end{aligned}$$

Thus $\text{bias}_f(\hat{f}_n(x)) = E_f\hat{f}_n(x) - f(x) = \int k(z)\{f(x-hz) - f(x)\}dz$. From this expression it is clear that the bias increases as h increases. By a standard Taylor expansion of $f(x - hz)$ about x we have

$$f(x - hz) = f(x) + f'(x)(-hz) + \frac{1}{2}f''(x^*)(-hz)^2$$

where $|x^* - x| \le h|z| \to 0$ if $h = h_n \to 0$. Then

$$\begin{aligned} h_n^{-2}\left\{E_f\hat{f}_n(x) - f(x)\right\} &= h_n^{-2}\int k(z)\{f'(x)(-h_n z) + \frac{1}{2}f''(x^*)(h_n z)^2\}dz \\ &= \frac{1}{2}\int f''(x^*)z^2 k(z)dz \to \frac{1}{2}f''(x)\int z^2 k(z)dz. \end{aligned}$$

By a Taylor expansion as in class, if $\int (f''(x))^2 dx < \infty$, it follows that

$$\int |\text{bias}_f(\hat{f}_n(x))|^2 dx \le h^4 \left(\int z^2 k(z)dz\right)^2 \int (f''(x))^2 dx \cdot (1/3).$$

(c) Using the fact that $\hat{f}_n$ is the sum of n independent and identically distributed terms

$$\begin{aligned} Var_f(\hat{f}_n(x)) &= \frac{1}{(nh)^2} n \cdot Var_f k\left(\frac{x - X_1}{h}\right) \\ &= \frac{1}{nh^2}\left\{E_f k^2\left(\frac{x - X_1}{h}\right) - \left\{E_f k\left(\frac{x - X_1}{h}\right)\right\}^2\right\} \\ &= \frac{1}{nh}\left\{\int \frac{1}{h} k^2\left(\frac{x - y}{h}\right) f(y)dy - h\left\{\int \frac{1}{h} k\left(\frac{x - y}{h}\right) f(y)dy\right\}^2\right\}, \end{aligned}$$

and hence if $h_n \to 0$,

$$\begin{aligned} nh_n Var_f(\hat{f}_n(x)) &= \int k^2(z) f(x - h_n z)dz - h_n\left\{\int k(z) f(x - h_n z)dz\right\}^2 \\ &\to f(x)\int k^2(z)dz. \end{aligned}$$

(d) Since the squared bias is of the order h_n^4 and the variance is of the order $1/(nh_n)$, the $h_n = O(n^{-1/5})$ gives the optimal order of the bandwidth. If we want the bias to satisfy

$$\sqrt{nh_n}(E_f(\hat{f}_n(x)) - f(x)) = \sqrt{nh_n} O(h_n^2)\sqrt{nh_n^5} \to 0,$$

then we would choose h_n to satisfy $nh_n^5 = o(1)$, or $h_n = o(n^{-1/5})$; e.g. if $h_n = n^{-\delta}$ with $1/5 < \delta < 1/2$. This is called *under smoothing*.

4. (40 points) The sample skewness $\hat{\beta}_{3,n}$ for data $X_1, \ldots, X_n$ is defined by

$$\hat{\beta}_{3,n} \equiv \frac{n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^3}{\{n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2\}^{3/2}}.$$

In the following assume that $X_1, \ldots, X_n$ are i.i.d. F with empirical distribution function $\mathbb{F}_n$.
(a) Write $\hat{\beta}_{n,3}$ explicitly as a functional $T(\mathbb{F}_n)$.
(b) If $E_F|X|^3 < \infty$ and $\sigma_F^2 \equiv Var_F(X_1) > 0$ do we have $T(\mathbb{F}_n) \to_p T(F)$ for some $T(F)$ described in terms of moments or central moments of F?
(c) If $F = N(\mu, \sigma^2)$ for some μ and σ^2, what is $T(F)$?
(d) If $E|X|^6 < \infty$, outline a proof of $\sqrt{n}(T(\mathbb{F}_n) - T(F)) \to_d N(0, V_F^2)$.
(No need to calculate V_F^2 explicitly.)
(e) Propose a resampling based estimator of the distribution

$$H_n(x, F) = P_F(\sqrt{n}(\hat{\beta}_{3,n} - \beta_3(F)) \leq)$$

not based on an explicit formula for V_F^2. How would you justify use of your proposed estimator?

Solution: (a) Since $(a-b)^3 = a^3 - 3a^2b + 3ab^2 - b^3$, the numerator can be written as

$$n^{-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^3 = \overline{X^3} - 3\overline{X}\cdot\overline{X^2} + 2\overline{X}^3$$

and the denominator can be written as $\{\overline{X^2} - \overline{X}^2\}^{3/2}$, it follows that

$$\hat{\beta}_{3,n} = g(\overline{X}, \overline{X^2}, \overline{X^3}) \equiv T(\mathbb{F}_n)$$

where $(\overline{X}, \overline{X^2}, \overline{X^3}) = \int (y, y^2, y^3) d\mathbb{F}_n(y)$ and where

$$g(x, y, z) = (z - 3xy + 2x^3)/(y - x^2)^{3/2}.$$

(b) If $E_F|X|^3 < \infty$, then

$$(\overline{X}, \overline{X^2}, \overline{X^3}) = \int (y, y^2, y^3) d\mathbb{F}_n(y) \to_{a.s.} \int (y, y^2, y^3) dF(y) = E_F(X, X^2, X^3).$$

Since g is continuous at points (x, y, z) with $y > x^2$, it follows by the continuous mapping theorem that

$$\begin{aligned} \hat{\beta}_{3,n} = g(\overline{X}, \overline{X^2}, \overline{X^3}) &\to_{a.s.} g(E_F(X), E_F(X^2), E_F(X^3) \\ &= E_F(X - \mu_F)^3/(\sigma_F^3) \equiv \beta_3(F) \equiv T(F). \end{aligned}$$

where $\mu_F \equiv E_F(X)$ and $\sigma_F^2 = Var_F(X)$.
(c) When $X \sim N(\mu, \sigma^2)$ then $\mu_F = \mu$, $\sigma_F^2 = \sigma^2$ and

$$T(F) = g(E_F(X), E_F X^2, E_F X^3) = E_F(X - \mu_F)^3/(\sigma_F^3) = EZ^3 = 0$$

where $Z \sim N(0, 1)$. Moreover, the same is true for any F symmetric about some μ with $E_F|Z|^3 < \infty$.
(d) When $E_F|X|^6 < \infty$, then, upon noting that we can, without loss, suppose that $\mu_F = 0$ and $\sigma_F^2 = 1$, it follows from the multivariate CLT that

$$\sqrt{n}\begin{pmatrix} \overline{X} \\ \overline{X^2} - 1 \\ \overline{X^3} - \beta_3 \end{pmatrix} \to_d \underline{Z} \sim N_3\left(0, \begin{pmatrix} 1 & \beta_3 & \gamma_2 + 3 \\ \beta_3 & \gamma_2 + 2 & \mu_5/\sigma^5 - \beta_3 \\ \gamma_2 + 3 & \mu_5/\sigma^5 - \beta_3 & \mu_6/\sigma^6 - \beta_3^2 \end{pmatrix}\right)$$

where $\gamma_2 \equiv EX^4/\sigma^4 - 3$ is the (excess of) kurtosis. Now the delta-method can be applied: since $\nabla g(0, 1, \beta_3) = (-3, -3\beta_3/2, 1)$ we conclude that

$$\sqrt{n}(\hat{\beta}_{3,n} - \beta_3(F)) \to_d \nabla g(0, 1, \beta_3)^T \underline{Z} \sim N_1(0, V^2(F))$$

for some $V^2(F)$. [When $F = N(0,1)$, $\beta_3 = 0$, $\gamma_2(F) = 0$, $\mu_5 = 0$, and $\mu_6 = 15$. The covariance matrix becomes

$$\Sigma_{N(0,1)} = \begin{pmatrix} 1 & 0 & 3 \\ 0 & 2 & 0 \\ 3 & 0 & 15 \end{pmatrix},$$

so $V^2(N(0,1)) = (-3,0,1)\Sigma(-3,0,1)^T = 15 - 2 \cdot 9 + 9 = 6$.]
Although this was not required, it is instructive to compute the influence function of $\beta_3(F)$ and compare with the above results. Thus we let $F_\epsilon = (1-\epsilon)F + \epsilon G$ and compute

$$\begin{aligned}
\frac{d}{d\epsilon}\beta_3(F_\epsilon)\Big|_{\epsilon=0} &= \frac{d}{d\epsilon}\int (x-\mu(F_\epsilon)^3 dF_\epsilon(x)/\sigma(F_\epsilon)^3\Big|_{\epsilon=0} \\
&= \frac{1}{\sigma_F^3}\int (x-\mu_F)^3 d(G-F) - \frac{3}{2}\frac{\mu_3(F)}{(\sigma_F^2)^{5/2}}\frac{d}{d\epsilon}\sigma(F_\epsilon)^2\Big|_{\epsilon=0} \\
&\qquad +\text{terms involving } \frac{d}{d\epsilon}\mu(F_\epsilon)\Big|_{\epsilon=0} \\
&= \int \left(\frac{x-\mu_F}{\sigma_F}\right)^3 d(G-F) - \frac{3}{2}\frac{\mu_3(F)}{(\sigma_F^2)^{3/2}}\cdot\int\left(\frac{x-\mu_F}{\sigma_F}\right)^2 d(G-F) \\
&\qquad +\text{ terms involving } \frac{d}{d\epsilon}\mu(F_\epsilon)\Big|_{\epsilon=0}.
\end{aligned}$$

Here the latter type of term coming from the numerator is

$$\begin{aligned}
&\frac{1}{\sigma_F^3}\int 3(x-\mu_F)^2 dF \cdot \left(-\frac{d}{d\epsilon}\mu(F_\epsilon)\Big|_{\epsilon=0}\right) \\
&\quad = -3\frac{\sigma_F^2}{\sigma_F^3}\int x d(G-F) = -3\int \frac{x-\mu_F}{\sigma} dG,
\end{aligned}$$

and the latter type of term coming from the denominator is

$$\frac{E_F(X-\mu_F)^3}{(\sigma_F^2)^{5/2}}(-3/2)\int 2(x-\mu_F)dF(x)\cdot\int x d(G-F) = 0.$$

Putting this all together yields

$$\begin{aligned}
\frac{d}{d\epsilon}\beta_3(F_\epsilon)\Big|_{\epsilon=0} &= \int \left(\frac{x-\mu_F}{\sigma_F}\right)^3 d(G-F) - \frac{3}{2}\frac{\mu_3(F)}{(\sigma_F^2)^{3/2}}\cdot\int\left(\frac{x-\mu_F}{\sigma_F}\right)^2 d(G-F) \\
&\qquad -3\int \frac{x-\mu}{\sigma_F} dG(x).
\end{aligned}$$

Thus, taking $G = \delta_x$ we find that the influence function for estimation of $\beta_3(F)$ is

$$\psi_F(x) = \left\{ \left(\frac{x - \mu_F}{\sigma_F} \right)^3 - \beta_3(F) \right\} - \frac{3}{2}\beta_3(F) \left\{ \left(\frac{x - \mu_F}{\sigma_F} \right)^2 - 1 \right\} - 3\frac{x - \mu_F}{\sigma_F}.$$

(e) The (ideal) bootstrap estimator of $H_n(x, F)$ is

$$H_n(x, \mathbb{F}_n) = P_{\mathbb{F}_n}(\sqrt{n}(T(\mathbb{F}_n^*) - T(F)) \leq x)$$

where T is the explicit formula for $\beta_3(F)$ from (a) and $\mathbb{F}_n^*$ denote the empirical distribution function of a bootstrap sample from $\mathbb{F}_n$. Since $\hat{\beta}_{3,n} = T(\mathbb{F}_n)$ is a Hadamard differentiable function of $\mathbb{F}_n$ our theory of the bootstrap for differentiable functionals implies that

$$\sup_x |H_n(x, \mathbb{F}_n) - H_n(x, F)| \to_p 0$$

since $\sqrt{n}(T(\mathbb{F}_n) - T(F)) \to_d N(0, V_F^2)$, and $\sqrt{n}(T(\mathbb{F}_n^*) - T(\mathbb{F}_n)) \to_d N(0, V_F^2)$ in probability.

Do either problem 5 or problem 6

5. (36 points). Let $T = T(F)$ be a (real-valued) function defined on a large class of distribution functions $\mathcal{F}$ which is large enough to contain all the empirical distribution functions. Suppose that $X_1, \ldots, X_n$ is a random sample from F, and let $T_n = T(\mathbb{F}_n)$ where $\mathbb{F}_n$ is the empirical distribution of the X_i's. Set $E_n = E_F(T_n) = E_F(T(\mathbb{F}_n))$, and suppose that we can write

$$E_n = T(F) + \frac{a_1(F)}{n} + \frac{a_2(F)}{n^2} + \cdots$$

so that the bias of $T_n = T(\mathbb{F}_n)$ is

$$\text{bias}_n(F) = E_F(T_n) - T(F) = \frac{a_1(F)}{n} + \frac{a_2(F)}{n^2} + \cdots .$$

(a) Describe the jack-knife estimator $\overline{T}_n^*$ of $T(F)$ and show that it has smaller bias than T_n.
(b) What is the ideal bootstrap estimator of $\text{bias}_n(F)$? Describe the corresponding bias corrected estimator $\overline{T}_{n,boot}^*$ of $T(F)$ via the bootstrap. Show that the resulting estimator has smaller bias when $T(F) = Var_F(X_1)$ and $T(\mathbb{F}_n) = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2$.

Solution: (a) For $i \in \{1, \ldots, n\}$, let $\mathbb{F}_{n,-i}$ denote the empirical distribution function of all the data with i left out: $\mathbb{F}_{n,-i}(x) = (n-1)^{-1}\sum_{j \neq i,} 1\{X_j \leq x\}$.

Let $T_{n,i} = T(\mathbb{F}_{n,-i})$, and $T_{n,\cdot} = n^{-1}\sum_1^n T_{n,i}$. Then $T^*_{n,i} = nT_n = (n-1)T_{n,i}$, $i = 1, \ldots, n$ are the pseudo-values. and $\overline{T}^*_n = n^{-1}\sum_{i=1}^n T^*_{n,i} = nT_n - (n-1)T_{n,\cdot}$; this is the jack-knife estimator of $T(F)$

Now with $T_n \equiv T(\mathbb{F}_n)$

$$E_n \equiv E_F T_n = T(F) + \frac{a_1(F)}{n} + \frac{a_2(F)}{n^2} + \cdots ,$$

we have

$$\begin{aligned} E_F(\overline{T}^*_n) &= nE(T_n) - (n-1)E_{n-1} \\ &= n\left\{T(F) + \frac{a_1(F)}{n} + \frac{a_2(F)}{n^2} + \cdots\right\} - (n-1)\left\{T(F) + \frac{a_1(F)}{n-1} + \frac{a_2(F)}{(n-1)^2} + \cdots\right\} \\ &= T(F) + a_2(F)\left\{\frac{1}{n} - \frac{1}{n-1}\right\} + \cdots \\ &= T(F) - \frac{1}{n-1}a_2(F) + \cdots . \end{aligned}$$

Thus the jack-knife estimator $\overline{T}^*_n$ has bias of order $O(n^{-2})$ whereas $T_n = T(\mathbb{F}_n)$ has bias of order $O(n^{-1})$.

(b) The bias is $\text{bias}_n(F) = E_F(T_n) - T(F)$. Hence the ideal bootstrap estimator of the bias is $\widehat{\text{bias}}_n \equiv E_{\mathbb{F}_n} T(\mathbb{F}^*_n) - T(\mathbb{F}_n)$ where $\mathbb{F}^*_n$ denotes the empirical distribution function of $X^*_1, \ldots, X^*_n$ i.i.d. from $\mathbb{F}_n$. When $T(F) = Var_F(X) = E_F(X^2) - (E_F(X))^2$, and hence

$$T(\mathbb{F}_n) = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2 = \overline{X^2} - \overline{X}^2 \equiv S^2_X$$

where $(\overline{X}, \overline{X^2}) = \int (x, x^2) d\mathbb{F}_n(x)$, we know that $T(\mathbb{F}_n)$ is a biased estimator: $E_F S^2_X = \frac{n-1}{n}\sigma^2_F$. Thus $(n/(n-1))S^2_X$ is an unbiased estimator, and the bias is just $\text{bias}_n(F) = -(1/n)\sigma^2_F$. Since this is true for a general F we also have

$$\widehat{\text{bias}}_n = E_{\mathbb{F}_n}(S^2_{X^*}) - T(\mathbb{F}_n) = \frac{n-1}{n}S^2_X - S^2_X = -(1/n)S^2_X.$$

This is clearly a consistent estimator of the bias, but it is, itself biased:

$$E_F\{-(1/n)S^2_X\} = -\frac{n-1}{n^2}\sigma^2_F.$$

Note that the bias corrected estimator is just

$$T(\mathbb{F}_n) - \widehat{\text{bias}}_n = S^2_X + \frac{1}{n}S^2_X = \frac{n+1}{n}S^2_X,$$

which has bias

$$\frac{(n+1)}{n} \cdot \frac{n-1}{n}\sigma_F^2 - \sigma_F^2 = \left\{ \frac{n^2-1}{n^2} - 1 \right\} \sigma_F^2 = -\frac{1}{n^2}\sigma_F^2.$$

Thus it appears that the jack-knife does somewhat better than the bootstrap in this case in that the bias of the jack-knife estimator is 0. [It turns out that there is an improvement to the bootstrap estimator of bias; see Efron and Tibshirani (1998), section 10.4, pages 130 ff.]

6. (36 points). Let $\underline{X}, \underline{X}_1, \ldots, \underline{X}_n$ be i.i.d. random vectors in $\mathbb{R}^d$ with $E\|\underline{X}\|^2 = E(\underline{X}_1^T \underline{X}) < \infty$. Conditionally on $\underline{X}_1, \ldots, \underline{X}_n$ with empirical measure $\mathbb{P}_n$, suppose that $\underline{X}^*_{n,1}, \ldots, \underline{X}^*_{n,n}$ are i.i.d. $\mathbb{P}_n$.
(a) Show that if $d = 1$, then for almost every sequence $X_1, X_2, \ldots$,

$$\sqrt{n}(\overline{X}^*_n - \overline{X}_n) \to_d N(0, Var(X)) \quad \text{as } n \to \infty.$$

In what sense does this show that "the bootstrap works"?
(b) Does the result in (a) continue to hold for $d > 1$? If yes, justify your answer.

Solution: We write

$$\sqrt{n}(\overline{X}^*_n - \overline{X}_n) = \sum_{i=1}^n Z_{ni}$$

where $Z_{ni} \equiv n^{-1/2}(X^*_{ni} - \overline{X}_n$ and proceed to apply the Lindeberg-Feller CLT:

$$E_* Z_{ni} = n^{-1/2}(E_* X^*_{ni} - \overline{X}_n) = n^{-1/2}(\overline{X}_n - \overline{X}_n) = 0,$$

$$\sigma^2_{ni} \equiv Var_*(Z_{ni}) = n^{-1}\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)^2, \quad \text{so that}$$

$$\sigma_n^2 = n^{-1}\sum_{i=1}^n (X_i - \overline{X}_n)^2 \to_{a.s.} \sigma_F^2.$$

It remains to verify the Lindeberg condition: to this end, we let $\epsilon > 0$ and compute

$$\begin{aligned}
&\frac{1}{\sigma_n^2}\sum_{i=1}^n E_*\{|Z_{ni}|^2 1_{[|Z_{ni}| \geq \epsilon \sigma_n]}\} \\
&= \frac{1}{\sigma_n^2} n \sum_{i=1}^n n^{-1}(X_i - \overline{X})^2 1_{[|X_i - \overline{X}_n| > \epsilon\sqrt{n}\sigma_n]} \\
&\leq \frac{1}{\sigma_n^2}\sum_{i=1}^n (X_i - \overline{X})^2 1_{[\max_{1\leq i\leq n}|X_i - \overline{X}_n| > \epsilon\sqrt{n}\sigma_n]} \\
&= 1_{[\max_{1\leq i\leq n}|X_i - \overline{X}_n| > \epsilon\sqrt{n}\sigma_n]} \to_{a.s.} 0
\end{aligned}$$

since $EX_1^2 < \infty$ implies

$$\begin{aligned} n^{-1} \max_{1\le i\le n} |X_i - \overline{X}_n|^2 &\le n^{-1} 2 \max_{1\le i\le n} |X_i - \mu|^2 + |\mu - \overline{X}_n|^2 \\ &\to_{a.s.} 0 \end{aligned}$$

Thus the LIndeberg-Feller CLT implies that for almost every sequence $X_1, X_2, \ldots$

$$\frac{\sqrt{n}(\overline{X}_n^* - \overline{X}_n)}{S_X} = \frac{\sum_{i=1}^n Z_{ni}}{\sigma_n} \to_d N(0,1).$$

Since $S_X^2 \to_{a.s.} \sigma_F^2 = Var_F(X_1)$, this implies the claimed conclusion: $\sqrt{n}(\overline{X}_n^* - \overline{X}_n) \to_d N(0, Var_F(X))$.
(b) Yes, it holds analogously in the In the case $d \ge 2$. We use the Cramér - Wold device. Let $a \in \mathbb{R}^d$. Then

$$\begin{aligned} a^T \sqrt{n}(\underline{\overline{X}}_n^* - \underline{\overline{X}}_n) &= \sqrt{n}(\overline{a^T \underline{X}}^* - \overline{a^T \underline{X}}_n) \\ &= \sqrt{n}(\overline{Y}_n^* - \overline{Y}_n) \end{aligned}$$

where $\overline{Y}_n = n^{-1} \sum_{i=1}^n a^T \underline{X}_i$ and $\overline{Y}_n^* = n^{-1} \sum_{i=1}^n a^T \underline{X}_i^*$. This converges in distribution for a.e. sequence $\underline{X}_1, \underline{X}_2, \ldots$ to $a^T N_d(0, \Sigma_F)$ by the result in (a) for $d = 1$. But by Cramér - Wold, this implies the desired convergence in distribution a.s. for the vectors $\sqrt{n}(\underline{\overline{X}}_n^* - \underline{\overline{X}}_n)$.

Do either Problem 7 or Problem 8.

7. (40 points) Suppose that $X_1, \ldots, X_n$ are i.i.d. with density f on $\mathbb{R}$. Let k be a probability density on $\mathbb{R}$ which is symmetric about zero and let $h > 0$. The local log-likelihood is, using Wasserman's notation,

$$\mathcal{L}_x(f) = \sum_{i=1}^n k\left(\frac{X_i - x}{h}\right) \log f(X_i) - n \int k\left(\frac{u-x}{h}\right) f(u) du.$$

If $f(u,a) = \exp(P_x(a,u))$ where $a = (a_0, \ldots, a_p) \in \mathbb{R}^{p+1}$ and

$$P_x(a,u) = a_0 + a_1(u-x) + \frac{1}{2} a_2 (u-x)^2 + \cdots + \frac{1}{p!} a_p (u-x)^p,$$

then maximizing $\mathcal{L}_x(f(u,a))$ over $a \in \mathbb{R}^{p+1}$ to find $\hat{a} = \text{argmax}_a \mathcal{L}_x(f(\cdot, a))$ yields a local polynomial smoothing estimator $\hat{f}_n(x) \equiv f(x, \hat{a})$ of f.
(a) Show that when $p = 0$, this $0-$th order polynomial smoothing estimator reduces to the kernel estimator of f based on k and the bandwidth h.
(b) When $p = 1$ find the equations satisfied by by $\hat{a} = (\hat{a}_0, \hat{a}_1)$ and investigate

the relationship between this new estimator and the standard kernel estimator obtained when $p = 0$.

Solution: (a) When $p = 0$, $f(u, a) = \exp(P_x(a, u)) = \exp(a_0)$, and hence

$$\begin{aligned}
\mathcal{L}_x(f(\cdot, a)) &= \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) \log f(X_i, a) - n \int k\left(\frac{u - x}{h}\right) f(u, a) du \\
&= \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) a_0 - n \int k\left(\frac{u - x}{h}\right) e^{a_0} du \\
&= a_0 \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) - n e^{a_0} \int k\left(\frac{u - x}{h}\right) du \\
&= a_0 \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) - nhe^{a_0} \equiv g(a_0)
\end{aligned}$$

since k is a density. To minimize this with respect to a_0 we compute

$$g'(a_0) = \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) - nhe^{a_0} = 0$$

if

$$e^{a_0} = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) = \frac{1}{nh} \sum_{i=1}^{n} k\left(x - \frac{X_i}{h}\right)$$

assuming that k is symmetric about 0. Thus the local likelihood density estimator with $p = 0$ is the classical kernel estimator.
(b) When $p = 1$, $f(u, a) = \exp(P_x(a, u)) = \exp(a_0 + a_1(u - x))$, and we find that, with $w_i(x) \equiv k((X_i - x)/h)$,

$$\begin{aligned}
\mathcal{L}_x(f(\cdot, a)) &= \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) \log f(X_i, a) - n \int k\left(\frac{u - x}{h}\right) f(u, a) du \\
&= \sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) (a_0 + a_1(X_i - x) - n \int k\left(\frac{u - x}{h}\right) e^{a_0 + a_1(u-x)} du \\
&= a_0 \sum_{i=1}^{n} w_i(x) + a_1 \sum_{i=1}^{n} w_i(x)(X_i - x) - n e^{a_0} \int k\left(\frac{u - x}{h}\right) e^{a_1(u-x)} du \\
&= a_0 \sum_{i=1}^{n} w_i(x) + a_1 \sum_{i=1}^{n} w_i(x)(X_i - x) - nhe^{a_0} \int k(z) \exp(ha_1 z) dz \\
&\equiv g(a_0, a_1)
\end{aligned}$$

Thus the (two) equations satisfied by $\hat{a}_0$ and $\hat{a}_1$ are given by

$$\begin{aligned}\frac{\partial g}{\partial a_0} &= \sum_{i=1}^{n} w_i(x) - nhe^{a_0}\int k(z)\exp(ha_1 z)dz = 0, \quad \text{and} \\ \frac{\partial g}{\partial a_0} &= \sum_{i=1}^{n} w_i(x)(X_i - x) - nh^2 e^{a_0}\int zk(z)\exp(ha_1 z)dz = 0.\end{aligned}$$

Solving the first equation for $e^{\hat{a}_0}$ yields

$$e^{\hat{a}_0} = \frac{\sum_{i=1}^{n} w_i(x)}{nh\int \exp(h\hat{a}_1 z)k(z)dz}, \tag{1}$$

where, from the second equation,

$$\sum_{i=1}^{n} w_i(x)(X_i - x) = nh^2\int \exp(h\hat{a}_1 z))zk(z)dz \cdot e^{\hat{a}_0}$$

Plugging (1) into the previous display and dividing by $\sum_1^n w_i(x)$ shows that we can rewrite the last display as

$$\frac{\sum_{i=1}^{n} w_i(x)(X_i - x)}{\sum_{i=1}^{n} w_i(x)} = h\frac{\int e^{h\hat{a}_1 z}zk(z)dz}{\int e^{h\hat{a}_1 z}k(z)dz}.$$

This shows that $h\hat{a}_1$ is the amount of tilting which makes the mean of the tilted distribution $e^{cz}k(z)/\int e^{cz'}k(z')dz'$ equal to the local mean

$$\frac{\sum_{i=1}^{n} w_i(x)((X_i - x)/h)}{\sum_{i=1}^{n} w_i(x)}.$$

Letting $\hat{a}_1 \to 0$ with fixed h gives back the $p = 0$ estimator as in (a) above. See Hjort and Jones (1996) for a thorough study of this and generalizations with $p \geq 2$.

8. (40 points) As noted in Wasserman (2006), one way (among many) of choosing the bandwidth h in density estimation problems is to minimize an appropriate estimator of the risk, $E_f L_2^2(f, \hat{f}_n(\cdot; h)) = E\left\{\int (\hat{f}_n(x; h) - f(x))^2 dx\right\}$. Noting that

$$\begin{aligned}L(h) &\equiv \int (\hat{f}_n(x; h) - f(x))^2 dx = \int \hat{f}_n^2(x; h) - 2\int \hat{f}_n(x; h)f(x)dx + \int f^2(x)dx \\ &\equiv J(h) + \int f^2(x)dx,\end{aligned}$$

where the last term does not depend on h, we can do this via a "leave-one-out" cross -validation estimator $\widehat{J}(h)$ of

$$J(h) = \int \hat{f}_n^2(x;h)dx - 2\int \hat{f}_n(x;h)f(x)dx$$

defined as follows:

$$\widehat{J}(h) \equiv \int \hat{f}_n^2(x;h)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{n,(-i)}(X_i,h)$$

where $\hat{f}_{n,(-i)}(x;h)$ is the estimator of f obtained in the same way as $\hat{f}_n$ by removing the $i-$th observation.
(a) Show that $E\hat{J}(h) = EJ(h)$ for kernel estimators $\hat{f}_n$ of f.
(b) Show that $E\hat{J}(h) = EJ(h)$ for histogram estimators $\hat{f}_n$ of f.

Solution: (From the solution set for problem set # 6) (a) (a) First the case of kernel density estimates: here

$$J(h) = \int \widehat{f}_n(x)^2dx - 2\int \widehat{f}_n(x)f(x)dx, \qquad \text{and}$$
$$\widehat{J}(h) = \int \widehat{f}_n(x)^2dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f}_{(-i)}(X_i),$$

so to show that $E\{\widehat{J}(h)\} = E\{J(h)\}$, it suffices to show that

$$E\left\{\frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{(-i)}(X_i)\right\} = E\left\{\int \widehat{f}_n(x)f(x)dx\right\}. \tag{2}$$

Now the right side in the last display is

$$\begin{aligned} E\left\{\int \widehat{f}_n(x)f(x)dx\right\} &= \int E\{\widehat{f}_n(x)\}f(x)dx \\ &= \iint \frac{1}{h}k\left(\frac{x-y}{h}\right)f(y)dyf(x)dx, \end{aligned}$$

while the left side is, by conditioning on X_i,

$$\begin{aligned} E\left\{\frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{(-i)}(X_i)\right\} &= \frac{1}{n}\sum_{i=1}^{n}E\left\{E(\widehat{f}_{(-i)}(X_i)|X_i)\right\} \\ &= \frac{1}{n}\sum_{i=1}^{n}E\{\int\frac{1}{h}k\left(\frac{X_i-x}{h}\right)f(x)dx\} \\ &= \iint\frac{1}{h}k\left(\frac{y-x}{h}\right)f(y)dyf(x)dx \\ &= \iint\frac{1}{h}k\left(\frac{x-y}{h}\right)f(y)dyf(x)dx \end{aligned}$$

by Fubini's theorem. Comparing the last two displays yields the claim.
(b) In the case of histogram estimators, it again suffices to show that (2) holds. But

$$\begin{aligned} E\left\{\int\widehat{f}_n(x)f(x)dx\right\} &= \int E\{\widehat{f}_n(x)\}f(x)dx = \int\sum_{j=1}^{m}\frac{p_j}{h}1_{B_j}(x)f(x)dx \\ &= \sum_{j=1}^{m}\frac{p_j^2}{h}, \end{aligned}$$

and, on the other hand,

$$\begin{aligned} E\left\{\frac{1}{n}\sum_{i=1}^{n}\widehat{f}_{(-i)}(X_i)\right\} &= E\left\{E\{n^{-1}\sum_{i=1}^{n}\widehat{f}_{(-i)}(X_i)|X_i\}\right\} \\ &= E\{n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{m}\frac{p_j}{h}1_{B_j}(X_i)\} \\ &= \sum_{j=1}^{m}\frac{p_j}{h}\int_{B_j}f(x)dx = \sum_{j=1}^{m}\frac{p_j^2}{h}, \end{aligned}$$

and hence we conclude that (again) $E\{\widehat{J}(h)\} = E\{J(h)\}$.