Econometrics V Lecture Notes (Spring 2016)

Tim Armstrong

last updated: February 17, 2016

1 Overview of topics

- Consider a statistical model where we observe $Y \overset{f}{\sim} P_f$, where f is known to be in a parameter space $\mathcal{F}$.
- Use P_f, E_f to denote probability and expectations under f
- We are interested in a functional Tf.
- Examples:
  - Gaussian regression model with fixed design:

    $$y_i = f(x_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2(x_i)), \quad \sigma^2(\cdot) \text{ known, } x_1, \ldots, x_n \text{ considered nonrandom}$$

    then

    $$Y \equiv \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \overset{f}{\sim} N\left(\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix}, \begin{pmatrix} \sigma^2(x_1) & \cdots & 0 \\ & \ddots & \\ 0 & \cdots & \sigma^2(x_n) \end{pmatrix} \right).$$

    - $*$ Examples of parameter spaces $\mathcal{F}$ for this model:
      - · Lipschitz: $\mathcal{F} = \mathcal{F}_{\text{Lip}}(C) = \{f : \mathbb{R} \to \mathbb{R} | |f(x) - f(x')| \leq C|x|\}$
      - · other smoothness classes: bounds on higher derivatives, other ways of bounding the derivative
      - · Linearity: $\mathcal{F} = \{x'\theta | \theta \in \mathbb{R}^{d_x}\}$
      - · Partial linear model $\{h(x_1) + x_2'\theta | \theta \in \mathbb{R}^{d_1}, h \in \widetilde{\mathcal{F}}\}$ for some $\widetilde{\mathcal{F}}$

        - Examples of functionals T in Gaussian regression model
            - $Tf = f(c)$ some fixed c
            - Regression discontinuity (RD): $Tf = \lim_{x\downarrow c} f(x) - \lim_{x\uparrow c} f(x)$
            - $Tf = \theta_k$ some fixed k in partial linear model
            - ATE under unconfoundedness: $Tf = \frac{1}{n}\sum_{i=1}^n [f(w_i, 1) - f(w_i, 0)]$ where $x_i = (w_i, d_i)$, d_i is indicator for "treatment"
        - In these examples, $\mathcal{F}$ is convex and T is linear
    - Nonparametric density estimation: x_i has density f wrt Lebesgue measure, interested in $Tf = f(c)$
        - Can use same choices for $\mathcal{F}$

- We are interested in a CI $\mathcal{C}$ for a functional Tf where $T : \mathcal{F} \to \mathbb{R}$ is a functional:

$$P_f(Tf \in \mathcal{C}) \geq 1 - \alpha \text{ all } f \in \mathcal{F} \tag{*}$$

  Subject to this constraint, we consider a performance criterion $R_f(\mathcal{C})$.

    - note that this depends on f
    - Examples
        - Expected length: $R_f(\mathcal{C}) = E_f \lambda(\mathcal{C})$ where λ denotes Lebesgue measure
        - Quantiles of excess length for one-sided CI: $R_f([\hat{c}, \infty)) = q_{f,\beta}(\hat{c} - Tf)$ where $q_{f,\beta}$ denotes the β quantile under f.

- Since R_f depends on f, we typically cannot choose $\mathcal{C}$ to minimize R_f simultatneously for all $f \in \mathcal{F}$.

- Efficiency bounds for CIs:

$$\text{minimize} \sup_{f \in \mathcal{G}} R_f(\mathcal{C}) \text{ s.t. } (*) \tag{**}$$

    - Setting $\mathcal{G} = \mathcal{F}$ gives minimax CI
    - Setting $\mathcal{G} \subsetneq \mathcal{F}$ gives sharp bound on adaptation

- Adaptive inference
    - How much does the value of (**) depend on $\mathcal{F}$?

  - Can $\mathcal{C}$ "adapt" to reflect the "smoothness" of $\mathcal{G}$ while maintaining coverage over $\mathcal{F}$ by making $\sup_{g\in\mathcal{G}} R_g(\mathcal{C})$ close to the minimax solution simultaneously for multiple $\mathcal{G}$?
  - Tells us whether and to what extent we can improve CIs through data driven choice of bandwidth, number of regressors, etc.

- Related problems that we won't cover or will not as spend much time on.
  - Adaptive/minimax estimation
  - Estimation/inference for the whole function: confidence bands for f, etc.

- Key ideas in solving our bounding (**)
  - Bounds by submodels: if we restrict $\mathcal{F}$ and $\mathcal{G}$ more, the problem can only become easier
  - Bounds by Bayesian problem: if we replace a worst-case criterion with an average (Bayesian) criterion, the problem can only become easier.
  - Use relation of CIs to hypothesis tests, bounds for optimal CIs ...
    - ... and use similar strategies to solve the optimal testing problem.

- Plan for this part of the course: start with optimal testing, then cover theory of minimax and adaptive inference with emphasis on linear functionals in Gaussian models

2 Minimax testing

- This section draws on parts of ch. 8 in Lehmann and Romano (2005) as well as Lehmann (1952), Section 2.4.3 of Ingster and Suslina (2003), and Emmanuel Candes' STAT300C lecture notes at Stanford.

- Observe $Y \sim P_f$, null hypothesis

$$H_0 : f \in \mathcal{F}_H$$

vs alternative

$$H_1 : f \in \mathcal{F}_K$$

- Level α test $\varphi(Y)$ satisfies

$$E_f\varphi(Y) \leq \alpha \text{ all } f \in \mathcal{F}_H.$$

- The test has minimax power (at least) β if

$$E_f\varphi(Y) \geq \beta \text{ all } f \in \mathcal{F}_K.$$

- Example: Suppose we want to determine whether the functional Tf is greater than some value T_0, where we know $f \in \mathcal{F}$. We can set $\mathcal{F}_H = \mathcal{F} \cap \{Tf \leq T_0\}$ and $\mathcal{F}_K = \mathcal{F} \cap \{Tf \geq T_0 + b\}$. Then, we ask how large we need b (the "effect size") to get a acceptable α and β (e.g. $\alpha = .05$, $\beta = .8$).
  - This can be a part of experimental design, or a part of an editor's decision of whether to accept or reject a paper.

- Suppose that P_f has density $p_f(y)$ wrt measure $\nu(y)$.

- Recall the Neyman-Pearson Lemma, which solves this problem for the case where $\mathcal{F}_H$ and $\mathcal{F}_K$ are singletons.

- Thm. (Neyman-Pearson Lemma): A most powerful test of $H_0 : f = f_0$ vs $H_1 : f = f_1$ exists and, for some c, satisfies

$$\varphi(y) = \begin{cases} 1 \text{ if } \frac{p_{f_1}(y)}{p_{f_0}(y)} > c \\ 0 \text{ if } \frac{p_{f_1}(y)}{p_{f_0}(y)} < c \end{cases}$$

  pf.: Omitted (see Thm. 3.2.1 in LR)

- Back to composite case. For a distribution Λ on $\mathcal{F}$, let

$$h_\Lambda(y) = \int_{f\in\mathcal{F}} p_f(y)d\Lambda(f).$$

- Let Λ_H and Λ_K be distributions on $\mathcal{F}_H$ and $\mathcal{F}_K$ respectively. Consider

$$H_{\Lambda_H} : Y \sim h_{\Lambda_H}(y) \text{ vs } H_{\Lambda_K} : Y \sim h_{\Lambda_K}(y).$$

- Lemma: Any level α test of H_0 vs H_1 with minimax power at least β is level α for H_{Λ_H} and has power at least β for H_{Λ_K}. In particular, no level α test of H_0 vs H_1 can have strictly greater minimax power than the (optimal) NP test of H_{Λ_H} vs H_{Λ_K}.

  pf.: The first statement follows since

$$\begin{aligned} E_{\Lambda_H}\varphi(Y) &= \int_y \varphi(y) h_{\Lambda_H}(y)\, d\nu(y) = \int_y \varphi(y) \int_{f\in\Lambda_H} p_f(y) d\Lambda_H(f)\, d\nu(y) \\ &= \int_{f\in\mathcal{F}_H} E_f\varphi(y)\, d\Lambda_H(f) \le \underbrace{\left[\sup_{f\in\mathcal{F}_H} E_f\varphi(y)\right]}_{\le\alpha} \underbrace{\int_{f\in\mathcal{F}_H} p_f(y)\, d\Lambda_H(f)}_{=1} \le \alpha \end{aligned}$$

  (using Fubini's theorem for the third equality) and similarly for minimax power. The last statement follows from this and optimality of the NP test for H_{Λ_H} vs H_{Λ_K}.

- Thm. (Theorem 8.1.1 in LR): Given Λ_H and Λ_K, let $\varphi_{\Lambda_H,\Lambda_K}$ be the NP test of h_{Λ_H} vs h_{Λ_K}, and let $\beta_{\Lambda_H,\Lambda_K}$ be its power at h_{Λ_K}. Suppose that there exist Λ_H and Λ_K such that

$$\sup_{f\in\mathcal{F}_H} E_f\varphi_{\Lambda_H,\Lambda_K}(Y) \le \alpha$$
$$\inf_{f\in\mathcal{F}_K} E_f\varphi_{\Lambda_H,\Lambda_K}(Y) = \beta_{\Lambda_H,\Lambda_K}$$

  (i.e. it is also level α for $\mathcal{F}_H$ with minimax power $\beta_{\Lambda_H,\Lambda_K}$ for $\mathcal{F}_K$). Then

  (i) $\varphi_{\Lambda_H,\Lambda_K}$ maximizes $\inf_{f\in\mathcal{F}_K} E_f\varphi_{\Lambda_H,\Lambda_K}(Y)$ among all level α tests of $\mathcal{F}_K$.

  (ii) The distributions Λ_H and Λ_K are least favorable in the sense that, for any other pair $\tilde{\Lambda}_H$, $\tilde{\Lambda}_K$, we have

$$\beta_{\Lambda_H,\Lambda_K} \le \beta_{\tilde{\Lambda}_H,\tilde{\Lambda}_K}.$$

  pf.: Part (i) is immediate from the lemma: any level α test with strictly greater minimax power would have power at h_{Λ_K} greater than $\beta_{\Lambda_H,\Lambda_K}$ and be level α for h_{Λ_H}, which would contradict optimality of $\varphi_{\Lambda_H,\Lambda_K}$ for h_{Λ_H} vs h_{Λ_K}.

  Part (ii) follows since, under the assumptions of the theorem, $\varphi_{\Lambda_H,\Lambda_K}$ is level α with minimax power $\beta_{\Lambda_H,\Lambda_K}$ for $\mathcal{F}_K$ and therefore (by the lemma) also level α for $h_{\tilde{\Lambda}_H}$ with power at least $\beta_{\Lambda_H,\Lambda_K}$ for $h_{\tilde{\Lambda}_H}$. Since this power is achievable, the optimal test, which achieves power $\beta_{\tilde{\Lambda}_H,\tilde{\Lambda}_K}$, must have greater power.

- Example (Lehmann, 1952): $Y \sim N(\theta, I_k)$, $\mathcal{F}_H = \{0\}$, $\mathcal{F}_K = \{\theta | \max_{1\leq j\leq k} |\theta_j| \geq b\}$.

- (draw picture)

- Consider

$$\Lambda_H = \text{unit mass at } 0$$
$$\Lambda_K(e_j b) = \Lambda_K(e_j(-b)) = \frac{1}{2k} \text{ all } j = 1, \ldots, k$$

where $e_j = (0, \ldots, 0, \underbrace{1}_{j\text{th position}}, 0, \ldots, 0)$ is the jth standard basis vector.

- NP test of H_{λ_H} vs H_{λ_K} rejects for large values of the LR:

$$\frac{\int \frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}(y-\theta)'(y-\theta)\right) \, d\Lambda_K(\theta)}{\frac{1}{(2\pi)^{k/2}} \exp\left(-\frac{1}{2}y'y\right)}$$
$$= \int \exp\left(y'\theta - \frac{1}{2}\theta'\theta\right) \, d\Lambda_K(\theta)$$
$$= \frac{1}{2k}\sum_{i=1}^{k}\left[\exp\left(by_j - \frac{1}{2}b^2\right) + \exp\left(-by_j - \frac{1}{2}b^2\right)\right].$$

- By the theorem, to show that this test is minimax, it suffices to show that its minimum power over $\mathcal{F}_K$ is equal to its average power under Λ_K. For this, it suffices to show that the power $E_\theta \phi_{\Lambda_H,\Lambda_K}$ is (i) minimized over $\theta \in \mathcal{F}_K$ at supp(Λ_K) and (ii) constant on supp(Λ_K).

- (ii) follows by symmetry. To show (i), note that, by symmetry, $E_{(\theta_1,\ldots,\theta_k)}\phi_{\Lambda_H,\Lambda_K} = E_{(|\theta_1|,\ldots,|\theta_k|)}\phi_{\Lambda_H,\Lambda_K}$, so we can restrict attention to the positive orthant. (i) will follow if we can show that $E_{(|\theta_1|,\ldots,|\theta_k|)}\phi_{\Lambda_H,\Lambda_K}$ is increasing in each element $|\theta_j|$, which will follow if the law of the LR statistic is increasing in each $|\theta_j|$ in the FOSD sense. This follows since the LR statistic is the sum of the k independent rvs,

$$\exp\left(bY_j - \frac{1}{2}b^2\right) + \exp\left(-bY_j - \frac{1}{2}b^2\right) \equiv W_j$$

where $Y_j \sim N(\theta_j, 1)$. W_j can be seen to be FOSD increasing in $|\theta_j|$ by noting that $P_{\theta_j}(W_j \leq t)$ is equal to the probability of Y_j being in a symmetric set around zero that depends only on t.

2.1 Bounds on Attainable Power

- How does the power of this test change with k and b?
- In general, difference between power and size of LR test of $H_0 : p_0$ vs $H_1 : p_1$ is

$$\begin{aligned} &E_{p_1} I\left(\frac{p_1(Y)}{p_0(Y)} \ge c_\alpha\right) - E_{p_0} I\left(\frac{p_1(Y)}{p_0(Y)} \ge c_\alpha\right) \\ &= E_{p_0}\left(\frac{p_1(Y)}{p_0(Y)} - 1\right) I\left(\frac{p_1(Y)}{p_0(Y)} \ge c_\alpha\right) \end{aligned}$$

- This is bounded by the <u>total variation distance</u>

$$\begin{aligned} TV(p_0, p_1) &\equiv E_{p_0}\left(\frac{p_1(Y)}{p_0(Y)} - 1\right) I\left(\frac{p_1(Y)}{p_0(Y)} \ge 1\right) \\ &= \int_{p_1(y) \ge p_0(y)} (p_1(y) - p_0(y))\, d\nu(y) \end{aligned}$$

- <u>Note</u>: This definition is symmetric since

$$\begin{aligned} &\int_{p_1(y) \ge p_0(y)} (p_1(y) - p_0(y))\, d\nu(y) + \int_{p_1(y) \le p_0(y)} (p_1(y) - p_0(y))\, d\nu(y) \\ &= \int (p_1(y) - p_0(y))\, d\nu(y) = 0 \end{aligned}$$

so that

$$\int_{p_1(y) \ge p_0(y)} (p_1(y) - p_0(y))\, d\nu(y) = \int_{p_1(y) \le p_0(y)} (p_0(y) - p_1(y))\, d\nu(y).$$

This also shows that

$$\begin{aligned} &TV(p_0, p_1) = \frac{1}{2}\left[\int_{p_1(y) \ge p_0(y)} (p_1(y) - p_0(y))\, d\nu(y) + \int_{p_1(y) \le p_0(y)} (p_0(y) - p_1(y))\, d\nu(y)\right] \\ &= \frac{1}{2} \int |p_1(y) - p_0(y)|\, d\nu(y). \end{aligned}$$

- One way of bounding the total variation distance is to use its equivalence with other distances such as the Hellinger distance (see Section 13.1 in Lehmann and Romano, 2005). We will not discuss this here.

- Lemma: Consider a sequence of densities $p_1^{(n)}(Y)$ and $p_0^{(n)}(Y)$ indexed by n. If

$$\frac{p_1^{(n)}(Y)}{p_0^{(n)}(Y)} \text{ converges in probability to 1 under } p_0^{(n)},$$

then $TV(p_0^{(n)}, p_1^{(n)}) \to 0$.

pf.: We have

$$TV(p_0^{(n)}, p_1^{(n)}) = E_{p_0^{(n)}} \left(\frac{p_1^{(n)}(Y)}{p_0^{(n)}(Y)} - 1 \right) I \left(\frac{p_1^{(n)}(Y)}{p_0^{(n)}(Y)} \geq 1 \right)$$
$$= -E_{p_0^{(n)}} \left(\frac{p_1^{(n)}(Y)}{p_0^{(n)}(Y)} - 1 \right) I \left(\frac{p_1^{(n)}(Y)}{p_0^{(n)}(Y)} \leq 1 \right)$$

using the fact that $E_{p_0^{(n)}} \left(\frac{p_1^{(n)}(Y)}{p_0^{(n)}(Y)} - 1 \right) = 0$. Since $\left(\frac{p_1^{(n)}(Y)}{p_0^{(n)}(Y)} - 1 \right) I \left(\frac{p_1^{(n)}(Y)}{p_0^{(n)}(Y)} \leq 1 \right)$ is bounded between -1 and 0, convergence in probability is sufficient for convergence of the expectation.

- Let us use this to see how large $b = b_k$ must be for power to go to zero.

- In the above example, LR is

$$\frac{h_{\Lambda_K}(Y)}{h_{\Lambda_H}(Y)} = \frac{1}{2} \left[\frac{1}{k} \sum_{i=1}^{k} \exp\left(bY_j - \frac{1}{2}b^2 \right) + \frac{1}{k} \sum_{i=1}^{k} \exp\left(-bY_j - \frac{1}{2}b^2 \right) \right].$$

- This is a sample average of k independent random variables with mean 1. We would like to use a LLN. The variance is increasing if b_k is increasing, so we need to use a triangular array argument.

- Consider each term individually. We have

$$var_0 \left(\exp\left(bY_j - \frac{1}{2}b^2 \right) \right) = E_0 \left[\exp\left(bY_j - \frac{1}{2}b^2 \right)^2 \right] - 1 = E_0 \exp\left(2bY_j - b^2 \right) - 1$$
$$= E_0 \exp\left((2b)Y_j - (2b)^2/2 + b^2 \right) - 1$$
$$= \exp(b^2) \underbrace{E \exp\left((2b)Y_j - (2b)^2/2 \right)}_{=1} - 1$$

so that

$$var_0\left(\frac{1}{k}\sum_{j=1}^{k}\exp\left(bY_j-\frac{1}{2}b^2\right)\right)=\frac{1}{k^2}\sum_{j=1}^{k}[\exp(b/2)-1]=\frac{1}{k}[\exp(b^2)-1].$$

If $b=b_k=C\sqrt{\log k}$, then this is $(\exp((C^2)\log k)-1)/k=(k^{C^2}-1)/k$, which converges to zero for $C<1$.

- We can improve this condition to $C<\sqrt{2}$ with a truncation argument. We will use some properties of the tail of the normal distribution.

- Lemma: Let ϕ and Φ denote the normal pdf and cdf. Then, for $t>0$,

$$1-\Phi(t)\leq\frac{\phi(t)}{t}.$$

pf.: We have

$$\begin{aligned}1-\Phi(t)&=\int_t^{\infty}\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)\,dz\leq\int_t^{\infty}\frac{z}{t}\frac{1}{\sqrt{2\pi}}\exp(-z^2/2)\,dz\\&=\frac{1}{t\sqrt{2\pi}}\left[-\exp(-z^2/2)\right]_{z=t}^{\infty}=\frac{1}{t\sqrt{2\pi}}\exp(-t^2/2).\end{aligned}$$

- Lemma: For Z_j iid $N(0,1)$,

$$P\left(\max_{1\leq j\leq k}Z_j\geq\sqrt{2\log k}\right)\to 0.$$

pf.: We have

$$\begin{aligned}&P\left(\max_{1\leq j\leq k}Z_j\geq\sqrt{2\log k}\right)\leq\sum_{j=1}^{k}P\left(Z_j\geq\sqrt{2\log k}\right)\\&=k[1-\Phi(\sqrt{2\log k})]\leq k\frac{1}{\sqrt{2\log k}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}(2\log k)\right)=(2\pi\cdot 2\log k)^{-1/2}\to 0.\end{aligned}$$

where the first step uses Bonferroni and the last step uses the previous lemma.

- Now we truncate the sum in the likelihood ratio.

- Let $X_j=\exp\left(bY_j-b^2/2\right)I\left(Y_j\leq\sqrt{2\log k}\right)$ and let W_j be defined in the same way with $-Y_j$ in place of Y_j. Then, by the above lemma,

$$\frac{h_{\Lambda_H}(Y)}{h_{\Lambda_K}(Y)} = \frac{1}{2}\left[\frac{1}{k}\sum_{j=1}^{k} X_j + \frac{1}{k}\sum_{j=1}^{k} W_j\right] \quad \text{w.p.a. 1.}$$

- Suffices to show that expectation of $\frac{1}{k}\sum_{j=1}^{k} X_j$ converges to 1 and variance converges to zero.

- We have

$$\begin{aligned} E_0 \frac{1}{k}\sum_{j=1}^{k} X_j = E_0 X_j &= \int_{-\infty}^{\sqrt{2\log k}} \exp\left(by - b^2/2\right) \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)\, dy \\ &= \int_{-\infty}^{\sqrt{2\log k}} \frac{1}{\sqrt{2\pi}} \exp(-(y-b)^2/2)\, dy \\ &= \Phi\left(\sqrt{2\log k} - b\right) \end{aligned}$$

  This converges to 1 for $b = C\sqrt{\log k}$ for $C < \sqrt{2}$.

- For the variance,

$$\begin{aligned} & var_0\left(\frac{1}{k}\sum_{j=1}^{k} X_j\right) = \frac{1}{k} var_0(X_1) \le \frac{1}{k} E_0 X_1^2 \\ &= \frac{1}{k}\int_{-\infty}^{\sqrt{2\log k}} \exp\left(2by - b^2\right) \frac{1}{\sqrt{2\pi}} \exp(-y^2/2)\, dy \\ &= \frac{1}{k}\exp(b^2) \int_{-\infty}^{\sqrt{2\log k}} \frac{1}{\sqrt{2\pi}} \exp(-(y-2b)^2/2)\, dy \\ &= \frac{1}{k}\exp(b^2)\Phi(\sqrt{2\log k} - 2b) \end{aligned}$$

- Set $C = (1-\varepsilon)\sqrt{2}$ so that $b = (1-\varepsilon)\sqrt{2\log k}$. Then the term in the $\Phi(\cdot)$ function is $\sqrt{2\log k} - 2(1-\varepsilon)\sqrt{2\log k} = (-1+2\varepsilon)\sqrt{2\log k}$, which gives

$$\begin{aligned} & \frac{1}{k}\exp((1-\varepsilon)^2 \cdot 2\log k)\Phi(-(1-2\varepsilon)\sqrt{2\log k}) \\ & \le \frac{1}{k}\exp(2(1-\varepsilon)^2\log k)\exp(-(1-2\varepsilon)^2\log k) = \exp(-2\varepsilon^2\log k) \end{aligned}$$

  using the bound $\Phi(-t) \le \exp(-t^2/2)$ for $t > 0$ (applies for ε small enough).

- Thus, minimax power goes to zero for $b = b_k = C\sqrt{\log k}$ for $C < \sqrt{2}$. How sharp is this bound?

- For $C > \sqrt{2}$, we can achieve this bound with a simpler test.

- Consider the test $\phi_{\text{Bonferroni}}(Y)$ defined by the rule

$$\text{reject if } \max_{1\le j\le k} |Y_j| > z_{1-\alpha/(2k)}$$

  where z_β denotes the β quantile of the $N(0,1)$ distribution.

- The test controls size by Bonferroni's inequality:

$$P_0\left(\max_{1\le j\le k} |Y_j| > z_{1-\alpha/(2k)}\right) \le \sum_{j=1}^{k} P\left(|Y_j| > z_{1-\alpha/(2k)}\right) = \sum_{j=1}^{k} \alpha/k = \alpha.$$

- We can obtain a bound on $z_{1-\alpha/(2k)}$:

$$\begin{aligned}
&1-\alpha/(2k) = \Phi(z_{1-\alpha/(2k)})\\
&\implies \alpha/(2k) = 1-\Phi(z_{1-\alpha/(2k)}) \le \exp(-z_{1-\alpha/(2k)}^2/2)\\
&\implies z_{1-\alpha/(2k)}^2/2 \le -\log(\alpha/(2k)) = \log k - \log(\alpha/2)\\
&\implies z_{1-\alpha/(2k)} \le \sqrt{2\log k - 2\log\alpha/2} = \sqrt{2\log k} + o(1).
\end{aligned}$$

- Thus, for θ with $\max_{1\le j\le k} |\theta_j| \ge C\sqrt{\log k}$, we have

$$E_\theta \phi_{\text{Bonferroni}}(Y) \ge E_{(C\sqrt{\log k},0,\dots,0)} \phi_{\text{Bonferroni}}(Y) \ge P_{(C\sqrt{\log k},0,\dots,0)}\left(|Y_1| > \sqrt{2\log k} + o(1)\right)$$
$$= P_{Z\sim N(0,1)}\left(\left|Z + C\sqrt{\log k}\right| \ge \sqrt{2\log k} + o(1)\right)$$

  which converges to one for $C > \sqrt{2}$.

- Summary:
  - Minimax power goes to one or α as $n \to \infty$ for $b = C\sqrt{\log k}$ depending on whether $C > \sqrt{2}$ or $C < \sqrt{2}$.
  - This is true regardless of the size α.
  - The Bonferroni test is "approximately minimax" as $k \to \infty$ in the sense that it also achieves power going to one for $C > \sqrt{2}$.

- An advantage of the Bonferroni test: it tells us which of the θ_j's are nonzero. See Chapter 9 of Lehmann and Romano (2005) and Donoho and Jin (2004) for more on these aspects of multiple testing.

2.2 Other notions of optimality

- def.: The power envelope at $g \in \mathcal{F} \backslash \mathcal{F}_H$ is given by the power of the MP test of $\mathcal{F}_H$ vs $\{g\}$.
- def.: The weighted average power (WAP) of the test φ for the weighting Λ is $\int E_f \varphi(Y)\, d\Lambda(Y)$.
- Note: By Fubini's theorem, WAP is

$$\int E_f \varphi(Y)\, d\Lambda(Y) = \int \int \varphi(y) p_f(y)\, d\nu(y) d\Lambda(f) = \int \varphi(y) \underbrace{\left[\int p_f(y)\, d\Lambda(f)\right]}_{h_\Lambda(y)} d\nu(y).$$

Thus, finding WAP optimal test reduces to testing $\mathcal{F}_H$ vs $\{h_\Lambda\}$.

2.3 Convex Hypotheses in the Normal Model

- The testing problem above is isomorphic to testing $f(x_i) = 0$ all i in the fixed design regression model. In this context, it might be too "pessimistic" to consider alternatives where $f(x_i) = b$ for just one observation i while being zero everywhere else.
- Solution: impose a priori restrictions - "smoothness." This often leads to problems that are related to testing convex hypotheses.
- Consider the model

$$Y = Kf + \varepsilon,$$

where $f \in \mathcal{F}$, K is known and

  - $\mathcal{F}$ (parameter space for f) is a convex subset of a vector space.
  - ε and Y take values in a Hilbert space $\mathcal{Y}$ with inner product $\langle \cdot, \cdot \rangle$, and ε is standard Gaussian with respect to this inner product: for any $g \in \mathcal{Y}$, $\langle g, \varepsilon \rangle \sim N(0, \|g\|^2)$.

  - $K : \mathcal{F} \to \mathcal{Y}$ is a (known) linear operator.

- This general setup follows Donoho (1994).

- In most examples in this course, $\mathcal{Y} = \mathbb{R}^n$ and $\langle x, y\rangle = x'y$. We won't worry about the details of defining Gaussian variables in infinite dimensional spaces, but see, e.g., Section 2.1 of Ingster and Suslina (2003) or Chapter 3 of Johnstone (2015) for details.

- Example (fixed design regression): $y_i = f(x_i) + u_i$, $u_i \sim N(0, \sigma^2(x_i))$ independent, $\sigma^2(x_i)$ known. We can set

$$\begin{aligned}
&Y = (y_1/\sigma(x_1), \ldots, y_n/\sigma(x_n)),\\
&Kf = (f(x_1)/\sigma(x_1), \ldots, f(x_n)/\sigma(x_n)),\\
&\mathcal{Y} = \mathbb{R}^n \text{ with } \langle x, y\rangle = x'y.
\end{aligned}$$

- Example (finite dimensional normal model): $W \sim N(\theta, \Sigma)$ where $\theta \in \mathbb{R}^k$, Σ known. We can set $Y = W$, $K = I$ and $\langle x, y\rangle = x'\Sigma^{-1}y$ (so that $\langle x, \varepsilon\rangle = x'\Sigma^{-1}\varepsilon \sim N(0, x'\Sigma^{-1}\Sigma\Sigma^{-1}x) = N(0, x'\Sigma^{-1}x)$ as required), or we can set $Y = \Sigma^{-1/2}W$, $K = \Sigma^{-1/2}$ and $\langle x, y\rangle = x'y$.

- Example (linear regression): $\underset{n\times 1}{Y} = \underset{n\times k}{X}\underset{k\times 1}{\theta} + \varepsilon$, $\varepsilon \sim N(0, I_n)$. Here, θ plays the role of f, and X (or the corresponding mapping from $\mathbb{R}^k$ to $\mathbb{R}^n$) plays the role of K.

- Example (Gaussian white noise): We observe $\{Y(t) | 0 \le t \le 1\}$ where

$$dY(t) = f(t) + dW(t)$$

  where $W(t)$ is a Brownian motion and $f \in L^2([0, 1])$. Heuristically, we can think of K as the identity and $dW(t)$ as the error term ε with "inner product" between g and dW given by $\int g(t)\, dW(t) \sim N(0, \|g\|^2)$. Linear estimators take the form $\int g(t)\, dY(t) \sim N(\langle g, f\rangle, \|g\|^2)$ for some $g \in L^2$ where $\langle f, g\rangle = \int f(t)g(t)\, dt$. Formally, we can identify f with its coefficients in an orthonormal basis $\{\psi_j(t)\}$ and consider the observation Y to be $\{\int \psi_j(t)\, dY(t)\}_{j=1}^{\infty}$. See p. 252 in Donoho (1994).

  - The white noise model is "asymptotically equivalent" to nonparametric density estimation and nonparametric regression under certain conditions (see Nussbaum, 1996; Brown and Low, 1996).

- Example (nonparametric IV with known first stage): Consider $y_i = h(w_i) + \eta_i$ with $E(\eta_i|z_i) = 0$. Then

$$E(y_i|z_i) = E(h(w_i)|z_i) = \int h(w)\, dF_{w|z}(w|z_i)$$

so we can write

$$y_i = \tilde{K}h(z_i) + \varepsilon_i$$

where $\tilde{K}$ takes the function h to the function $z \mapsto \int h(w)\, dF_{w|z}(w|z)$. If we condition on the z_i's and assume the distribution of $w|z$ is known, then this fits into our framework with K mapping h to the coordinates of $\tilde{K}h$. Typically, however, $\tilde{K}$ is not known.

  – Related inverse problems such as deconvolution have been put into this framework. See, e.g., examples in Donoho and Low (1992).

- We will use the following fact about the Gaussian shift model, which we state as a proposition.

- Proposition: Consider $Y = \theta + \varepsilon$, $\theta \in \mathcal{Y}$ (i.e. θ plays the role of Kf), ε standard Gaussian. The likelihood ratio statistic for the simple testing problem $H_0 : \theta_0$ vs $H_1 : \theta_1$ is given by

$$\frac{p_1(Y)}{p_0(Y)} = \exp\left(\langle Y, \theta_1 - \theta_0\rangle - \frac{1}{2}\|\theta_1 - \theta_0\|^2\right).$$

In particular, the Neyman-Pearson test rejects for large values of $\langle Y, \theta_1 - \theta_0\rangle$. The power of the level α Neyman-Pearson test is

$$\Phi(\|\theta_1 - \theta_0\| - z_{1-\alpha}).$$

pf.: The first claim follows by simple calculations in the finite dimensional case. For the infinite dimensional case, see, e.g. Chapter 3 of Johnstone (2015) for details (in the sequence model). For the power of the test, note that

$$\langle Y, \theta_1 - \theta_0\rangle \overset{\theta}{\sim} N(\langle \theta, \theta_1 - \theta_0\rangle, \|\theta_1 - \theta_0\|^2)$$

so the test rejects when

$$\begin{aligned}&\langle Y, \theta_1 - \theta_0\rangle - \langle \theta_0, \theta_1 - \theta_0\rangle > \|\theta_1 - \theta_0\| z_{1-\alpha} \\ \iff &\frac{\langle Y - \theta_0, \theta_1 - \theta_0\rangle}{\|\theta_1 - \theta_0\|} > z_{1-\alpha}.\end{aligned}$$

Under θ_1, $\frac{\langle Y-\theta_0, \theta_1-\theta_0\rangle}{\|\theta_1-\theta_0\|} \overset{\theta_1}{\sim} N(\|\theta_1 - \theta_0\|, 1)$, so the power is

$$P_{\theta_1}\left(\frac{\langle Y - \theta_0, \theta_1 - \theta_0\rangle}{\|\theta_1 - \theta_0\|} > z_{1-\alpha}\right) = P_{Z\sim N(0,1)}\left(Z + \|\theta_1 - \theta_0\| > z_{1-\alpha}\right) = \Phi\left(\|\theta_1 - \theta_0\| - z_{1-\alpha}\right).$$

- Consider observing $Y = Kf + \varepsilon$ (satisfying the Gaussian model above) and testing $H_0 : f \in \mathcal{F}_H$ vs $H_0 : f \in \mathcal{F}_K$, where $\mathcal{F}_H$ and $\mathcal{F}_K$ are convex.
- Strategy for finding optimal test: conjecture that the least favorable distributions are point masses.
- Bound on minimax power from two-point testing problem f vs g for $f \in \mathcal{F}_H$, $g \in \mathcal{F}_K$:

  $$\Phi\left(\|K(f - g)\| - z_{1-\alpha}\right)$$

  (in the notation of the general minimax testing theorem, this is $\beta_{\Lambda_f,\Lambda_g}$ where Λ_f and Λ_g are point masses at f and g respectively).
- By the theorem, f and g must minimize this, which is equivalent to

  $$\min_{f,g} \|K(f - g)\| \text{ s.t. } f \in \mathcal{F}_H, g \in \mathcal{F}_K. \qquad (*)$$

- Thm. (see Section 2.4.3 in Ingster and Suslina (2003)): Let (f^*, g^*) be a pair that minimizes (*) (assume that the minimum is achieved).
  - (i) The level α NP test of f^* vs g^* is the minimax optimal test for $\mathcal{F}_H$ vs $\mathcal{F}_K$ at level α.
  - (ii) The minimax power is $\Phi(\|K(f^* - g^*)\| - z_{1-\alpha})$.
  - (iii) The rejection probability is maximized over $\mathcal{F}_H$ at f^* and minimized over $\mathcal{F}_K$ at g^*.

pf.: By the general minimax theorem, (i) and (ii) will follow if we can show (iii). By the calculations above, the NP test ϕ of f^* vs g^* rejects for large values of $\langle Y, K(f^* - g^*)\rangle$ with constant critical value. Since

$$\langle Y, K(f^* - g^*)\rangle \overset{f}{\sim} N\left(\langle Kf, K(f^* - g^*)\rangle, \|K(f^* - g^*)\|^2\right),$$

the rejection probability $E_f\phi$ is an increasing function of $\langle Kf, K(f^* - g^*)\rangle$. Thus, it suffices to show that this is maximized over $\mathcal{F}_H$ at f^* and minimized over $\mathcal{F}_K$ at g^*.

To show the latter (the former is symmetric), let $g \in \mathcal{F}_K$, and let $g_\lambda = g\lambda + g^*(1-\lambda) = g^* + \lambda(g - g^*)$. By convexity, $g_\lambda \in \mathcal{F}_K$ for $\lambda \in [0,1]$, so, by optimality of g^*, we must have $\|K(f^* - g_\lambda)\|^2$ minimized at $\lambda = 0$ over $\lambda \in [0,1]$. Thus, the derivative at zero is nonnegative:

$$\begin{aligned}
0 &\le \frac{d}{d\lambda_+}\|K(f^* - g_\lambda)\|^2\bigg|_{\lambda=0} \\
&= \frac{d}{d\lambda_+}\|K(f^* - g^*)\|^2 + 2\lambda\langle K(g - g^*), K(f^* - g^*)\rangle + \lambda^2\|K(g - g^*)\|^2\bigg|_{\lambda=0} \\
&= 2\langle K(g - g^*), K(f^* - g^*)\rangle
\end{aligned}$$

so $\langle Kg, K(f^* - g^*)\rangle$ is minimized over $\mathcal{F}_K$ at g^* as required.

- Example (one-sided testing for a linear functional): Let $L : \mathcal{F} \to \mathbb{R}$, and consider the minimax testing problem

  $$H_0 : Lf \le L_0 \text{ and } f \in \mathcal{F} \text{ vs } H_1 : Lf \ge L_0 + b \text{ and } f \in \mathcal{G}$$

  where $\mathcal{F}$ and $\mathcal{G}$ are convex sets. Since $\{f | Lf \le L_0\} \cap \mathcal{F}$ and $\{f | Lf \ge L_0 + b\} \cap \mathcal{G}$ are convex sets, the theorem applies. Setting $\mathcal{G} = \mathcal{F}$ gives a minimax criterion, setting $\mathcal{G} \subsetneq \mathcal{F}$ gives exact bound on adaptation.

  As a special case, consider the fixed design regression model with

  $$Lf = f(0) \qquad \text{and} \qquad \mathcal{F} = \mathcal{F}_{\text{Lip}}(C) = \{f : \mathbb{R} \to \mathbb{R} | |f(x) - f(x')| \le C|x - x'|\}.$$

For simplicity, suppose $\sigma^2(x) = 1$. Then f^*, g^* solves

$$\min_{f,g} \sum_{i=1}^{n} (f^*(x_i) - g^*(x_i))^2 \text{ s.t. } f(0) \le L_0,\ g(0) \ge L_0 + b,\ f \in \mathcal{F},\ g \in \mathcal{G}.$$

- For the minimax test ($\mathcal{F} = \mathcal{G}$),

  $$f^*(x) = L_0 + b/2 - \max\{b/2 - C|x|, 0\}, \quad g^*(x) = L_0 + b/2 + \max\{b/2 - C|x|, 0\}$$

  (draw picture).

- Test statistic:

  $$\sum_{i=1}^{n} (g^*(x_i) - f^*(x_i))y_i = \sum_{i=1}^{n} 2\max\{b/2 - C|x_i|, 0\}y_i = b\sum_{i=1}^{n} k(x_i/h)y_i$$

  where $k(x) = \max\{1 - |x|, 0\}$ and $h = b/(2C)$. Critical value is $1-\alpha$ quantile under f^*:

  $$\begin{aligned}&\sum_{i=1}^{n} (g^*(x_i) - f^*(x_i))f^*(x_i) + z_{1-\alpha}\sqrt{\sum_{i=1}^{n} (g^*(x_i) - f^*(x_i))^2} \\ &= b\sum_{i=1}^{n} k(x_i/h)(L_0 + b/2 - (b/2)k(x_i/h)) + z_{1-\alpha}b\sqrt{\sum_{i=1}^{n} k(x_i/h)^2}\end{aligned}$$

  Rearranging, the rejection region is

  $$\frac{\sum_{i=1}^{n} k(x_i/h)y_i}{\sum_{i=1}^{n} k(x_i/h)} > L_0 + (b/2)\left(1 - \frac{\sum_{i=1}^{n} k(x_i/h)^2}{\sum_{i=1}^{n} k(x_i/h)}\right) + z_{1-\alpha}\frac{\sqrt{\sum_{i=1}^{n} k(x_i/h)^2}}{\sum_{i=1}^{n} k(x_i/h)}.$$

  Inverting the tests leads to the CI $[\hat{c}, \infty)$ where

  $$\hat{c} = \hat{L}_h - \overline{\text{bias}}(\hat{L}_h) - z_{1-\alpha}\text{se}(\hat{L}_h)$$

  with $\hat{L}_h = \frac{\sum_{i=1}^{n} k(x_i/h)y_i}{\sum_{i=1}^{n} k(x_i/h)}$, $\overline{\text{bias}}(\hat{L}_h) = (b/2)\left(1 - \frac{\sum_{i=1} k(x_i/h)^2}{\sum_{i=1}^{n} k(x_i/h)}\right)$ and $\text{se}(\hat{L}_h) = \frac{\sqrt{\sum_{i=1}^{n} k(x_i/h)^2}}{\sum_{i=1}^{n} k(x_i/h)}$.

- Now consider directing power at constant functions: $\mathcal{G} = \{f(x) : f(x) = c \text{ some } c \in \mathbb{R}\}$, and with distance to null given by $b/2$ instead of b. Then f^* is the same as before and $g^*(x) = L_0 + b/2$. Test statistic

$$\sum_{i=1}^{n}(g^*(x_i) - f^*(x_i))y_i = \sum_{i=1}^{n}\max\{b/2 - C|x_i|, 0\}y_i = (b/2)\sum_{i=1}^{n}k(x_i/h)y_i$$

proportional to minimax (for original b) test statistic and critical values based on same f^*, so this leads to same test as before.

- Can also calculate power of minimax test at constant functions: it is typically close to the power envelope.
- We will develop a comprehensive theory that covers this example and other adaptivity bounds later in the course.

- Example (power envelope for moment inequalities, c.f. Romano, Shaikh, and Wolf 2013): Consider the fixed-design regression model. The null hypothesis

$$H_0 : f(x) \leq 0 \text{ all } x$$

is a conditional moment inequality. The power envelope at g is given by the power of the test of H_0 vs $H_1 : \{g\}$. Since H_0 and H_1 are convex, the result applies.

  - Often, we are interested in the parameter θ where we use

$$f(x) = E(m(W_i, \theta_0)|X_i = x)$$

  to test the null $\theta_0 \in \Theta_I$, where

$$\Theta_I = \{\theta | E(m(W_i, \theta_0)|X_i = x) \text{ all } x\}$$

  is the identified set. Inverting these tests, we can get a CI for θ. In this setting, we are interested in minimax relative efficiency of CIs for θ or Θ_I. See Armstrong (2014).

2.4 Unions of convex hypotheses

- Often, nonconvex hypotheses can be written as unions of convex hypotheses. Then, one can often use ad hoc strategies for bounding the minimax power and deriving "approximately" minimax tests based on the individual pairs of least favorable functions. We will give some examples, which we will come back to later in the course.

- Example (minimax one-sided testing in the sup norm): Consider a version of the sup-norm testing problem in the fixed design regression model. For concreteness, suppose that $f : [0, 1] \to \mathbb{R}$ (so that $x_i \in [0, 1]$) and let

$$\begin{aligned} &H_0 : f = 0 \\ &H_1 : \max_{1 \le i \le n} f(x_i) \ge b \text{ and } f \in \mathcal{F}. \end{aligned}$$

  If $\mathcal{F}$ is unrestricted, this is equivalent to a one-sided version of the original problem. H_1 is not convex, but we can write it as

$$H_1 : \cup_{i=1}^{n} \{ f \in \mathcal{F} | f(x_i) \ge b \}$$

  which is the union of n convex alternatives.

- Example (adaptive testing for linear functionals): consider testing

$$H_0 : Lf \le L_0 \text{ and } f \in \mathcal{F} \text{ vs } H_1 : \text{there exists } C \in \mathcal{I} \text{ s.t. } Lf \ge L_0 + b(C) \text{ and } f \in \mathcal{G}(C)$$

  where $b(C)$ is some function of C, $\mathcal{G}(C)$ is convex for each C and $\mathcal{I}$ is some index set. For example, we can take $\mathcal{G}(C) = \mathcal{F}_{\text{Lip}}(C)$ and $\mathcal{I}$ to be some real interval. This is a problem of adaptive testing. Interest often focuses on asymptotic results such as finding $\tilde{b}(C, n)$ and K^*, K_* such that minimax power goes to one for $b(C) = K^* \tilde{b}(C, n)$ and zero for $b(C) = K_* \tilde{b}(C, n)$.

2.5 Brute force computational approach to optimal testing

- According to the optimal testing theorem, the optimal test can be found by solving the problem

$$\min \beta_{\Lambda_H, \Lambda_K} \quad \text{s.t.} \quad \Lambda_H \text{ a distribution on } \mathcal{F}_H, \quad \Lambda_K \text{ a distribution on } \mathcal{F}_K.$$

- Suppose that $\mathcal{F} = \mathbb{R}^k$. Then we can approximate this with the solution to

  $$\min \beta_{\Lambda_H,\Lambda_K} \quad \text{s.t.} \quad \Lambda_H \text{ a distribution on } \varepsilon\mathbb{Z}^k \cap \mathcal{F}_H, \quad \Lambda_K \text{ a distribution on } \varepsilon\mathbb{Z}^k \cap \mathcal{F}_K.$$

  For any ε, this gives an upper bound.

- Can also approximate $\mathcal{F}_H$ in other ways (e.g. basis functions instead of grid).

- Main problem: computational curse of dimensionality - number of grid points (or basis functions, etc.) grows exponentially with k. In the fixed design regression model, $k = n = \#$ of observations.

- See Elliott, Müller, and Watson (2015) for recent applications (using WAP instead of minimax) and references therein for more detail.

3 CIs and estimation for linear functionals

- This section covers parts of Donoho (1994), Cai and Low (2004) and Armstrong and Kolesár (2015).

- Consider inference on a linear functional Lf in the general Gaussian model

  $$y = Kf + \varepsilon.$$

- We expect that optimal (minimax/adaptive) CIs are related to the testing problem

  $$H_0 : Lf \leq L_0 \text{ and } f \in \mathcal{F} \text{ vs } H_0 : Lf \geq L_0 + b \text{ and } g \in \mathcal{G}$$

  We showed that the minimax test was the NP test of f^* vs g^* where (f^*, g^*) minimize $\|K(g-f)\|$ subject to $f \in H_0$ and $g \in H_1$.

- If we minimize over L_0 as well, we get the problem

  $$\min \|K(f-g)\| \text{ s.t. } Lg - Lf \geq b, f \in \mathcal{F}, g \in \mathcal{G}. \tag{*}$$

  The dual of this problem is

  $$\max Lg - Lf \text{ s.t. } \|K(f-g)\| \leq \delta, f \in \mathcal{F}, g \in \mathcal{G}. \tag{**}$$

- def.: The maximized value of (**) is called the (ordered) modulus of continuity, and is denoted

$$\omega(\delta) = \omega(\delta; \mathcal{F}, \mathcal{G}) = \omega(\delta; \mathcal{F}, \mathcal{G}, K, L).$$

The minimized value of (*) is called the inverse (ordered) modulus of continuity, and is denoted

$$\omega^{-1}(b) = \omega^{-1}(b; \mathcal{F}, \mathcal{G}) = \omega^{-1}(b; \mathcal{F}, \mathcal{G}, K, L).$$

- Note: Using the characterization of the two-point NP test, it follows that $\omega(\delta)$ maximizes $Lf - Lg$ subject to the constraint that the NP test has power less than $\Phi(\delta - z_{1-\alpha})$.

- Thm (properties of the modulus): Assume that $\mathcal{F} \cap \mathcal{G} \neq \emptyset$. Then

  1.) $\omega(\delta)$ is nonnegative, nondecreasing and concave.

  2.) For some $\overline{\delta}$ (possibly ∞ or 0), $\omega(\delta)$ is strictly increasing on $[0, \overline{\delta})$ and constant on $[\overline{\delta}, \infty)$. Its inverse $\omega^{-1} : [\underline{b}, \overline{b}) \to [0, \overline{\delta})$ (where $\underline{b} = \omega(0)$ and $\overline{b} = \lim_{\delta \to \overline{\delta}} \omega(\delta)$) is given by the solution to (*).

pf.: Nonnegativity follows since $f = g$ is feasible. The modulus is increasing since the constraint set increases with δ. For concavity, note that, if f^*_δ, g^*_δ attain the modulus at δ and $f^*_{\tilde{\delta}}, g^*_{\tilde{\delta}}$ attain the modulus at $\tilde{\delta}$, then, for $\lambda \in [0, 1]$, $g_\lambda = \lambda g^*_\delta + (1 - \lambda) g^*_{\tilde{\delta}}$ and $f_\lambda = \lambda f^*_\delta + (1 - \lambda) f^*_{\tilde{\delta}}$ satisfy $\|K(g_\lambda - f_\lambda)\| \leq \lambda\delta + (1 - \lambda)\tilde{\delta}$, so that $\omega(\lambda\delta + (1 - \lambda)\tilde{\delta}) \geq L(g_\lambda - f_\lambda) = \lambda\omega(\delta) + (1 - \lambda)\omega(\tilde{\delta})$ (if the moduli are not achieved, we can argue similarly with limits).

The first part of (2) follows since a nondecreasing concave function cannot be strictly increasing only on disjoint sets. To show that the inverse is the solution to (*), we need to show that solving (*) with $b = \omega(\delta)$ gives a minimized value of δ for $\delta < \overline{\delta}$. If this minimized value were strictly less than δ, we would have $\|K(f - g)\| < \delta$ for some f, g with $Lg - Lf \geq \omega(\delta)$. Then, for some $\delta' > \delta$ with $\omega(\tilde{\delta}) > \omega(\delta)$, we could strictly increase $Lg - Lf$ by taking a convex combination with a pair $f_{\tilde{\delta}}, g_{\tilde{\delta}}$ that is within a small enough constant of achieving the modulus at $\tilde{\delta}$. If, the minimized value of (*) were strictly greater than δ, then there would exist $\eta > 0$ such that $\|K(f - g)\| > \delta + \eta$ for all f, g with $Lg - Lf \geq \omega(\delta)$. This would imply $\omega(\delta + \eta) \leq \omega(\delta)$.

- By concavity, ω has a nonempty superdifferential

$$\partial\omega(\delta) \equiv \{d | \text{for all } \eta > 0,\, \omega(\eta) \leq \omega(\delta) + d(\eta - \delta)\}$$

  on $(0, \infty)$.

- If ω is differentiable (which is typically the case), then $\partial\omega(\delta)$ is the singleton set containing the derivative at δ.

- Lemma: Let f^* and g^* achieve the modulus at δ with $\|K(f^* - g^*)\| = \delta$. Then, for any $d \in \partial\omega(\delta)$ in the superdifferential, we have, for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,

$$L(g - g^*) \leq d\frac{\langle K(g^* - f^*), K(g - g^*)\rangle}{\|K(g^* - f^*)\|} \text{ and } L(f - f^*) \geq d\frac{\langle K(g^* - f^*), K(f - f^*)\rangle}{\|K(g^* - f^*)\|}.$$

  pf.: We will prove the first inequality (the second is symmetric). Given $g \in \mathcal{G}$, let $g_\lambda = \lambda g + (1 - \lambda)g^*$. By convexity, $g_\lambda \in \mathcal{G}$. Note that

$$g_\lambda - f^* = \lambda g + (1 - \lambda)g^* - f^* = \lambda(g - g^*) + g^* - f^*$$

  so that

$$L(g_\lambda - f^*) = \lambda L(g - g^*) + L(g^* - f^*) = \lambda L(g - g^*) + \omega(\delta)$$

  and

$$\frac{d}{d\lambda_+}\|K(g_\lambda - f^*)\|\bigg|_{\lambda=0} = \frac{1}{2}\frac{\frac{d}{d\lambda_+}\|K(g_\lambda - f^*)\|^2\Big|_{\lambda=0}}{\|K(g^* - f^*)\|} = \underbrace{\frac{\langle K(f^* - g^*), K(g - g^*)\rangle}{\|K(g^* - f^*)\|}}_{\equiv\eta}$$

  (last equality uses same calculations as in convex testing lemma). Thus,

$$\begin{aligned}
&\lambda L(g - g^*) = L(g_\lambda - f^*) - \omega(\delta) \\
&\leq \omega(\|K(g_\lambda - f^*)\|) - \omega(\delta) && \text{(by def. of modulus)} \\
&\leq d\,[\|K(g_\lambda - f^*)\| - \delta] && \text{(by def. of } d\text{)} \\
&= d(\lambda\eta + o(\lambda)) && \text{(since } \eta \text{ is the derivative at 0).}
\end{aligned}$$

Dividing by both sides by λ and taking limit as $\lambda \to 0$ gives the result.

- We can use the lemma to construct an estimate based on NP test for least favorable pair.

- Let $\omega'(\delta)$ denote an arbitrary element in $\partial\omega(\delta)$ (i.e. $\omega'(\delta) = d$ in the notation of the previous lemma). Let f^*_δ, g^*_δ denote the pair that achieves the modulus at δ (we will sometimes drop the subscript when it is clear).

- Consider the statistic

$$T_\delta = \frac{\omega'(\delta)}{\delta}\langle K(g^*_\delta - f^*_\delta), Y\rangle.$$

For $g \in \mathcal{G}$,

$$\begin{aligned}\text{bias}_g(T_\delta) \equiv E_g T_\delta - Lg = \frac{\omega'(\delta)}{\delta}\langle K(g^*_\delta - f^*_\delta), Kg\rangle - Lg \\ \geq \frac{\omega'(\delta)}{\delta}\langle K(g^*_\delta - f^*_\delta), Kg^*_\delta\rangle - Lg^*_\delta = \text{bias}_{g^*_\delta}(T_\delta),\end{aligned}$$

where the inequality follows from the lemma. Similarly, for $f \in \mathcal{F}$,

$$\text{bias}_f(T_\delta) \leq \text{bias}_{f^*_\delta}(T_\delta).$$

- Thus,

$$T_\delta \overset{f}{\sim} N\left(Lf + \text{bias}_f(T_\delta), [\omega'(\delta)]^2\right)$$

where,

$$\begin{aligned}\text{bias}_f(T_\delta) \leq \text{bias}_{f^*_\delta}(T_\delta) \text{ for } f \in \mathcal{F} \\ \text{bias}_g(T_\delta) \geq \text{bias}_{g^*_\delta}(T_\delta) \text{ for } g \in \mathcal{G}.\end{aligned}$$

- Lemma: Suppose that the modulus is achieved at δ with $\|K(f_\delta^* - g_\delta^*)\| = \delta$. Let

$$\begin{aligned} B &= \text{bias}_{f_\delta^*}(T_\delta) - \text{bias}_{g_\delta^*}(T_\delta) \\ &= \frac{\omega'(\delta)}{\delta} \underbrace{\langle K(g_\delta^* - f_\delta^*), K(f_\delta^* - g_\delta^*)\rangle}_{=-\delta^2} - \underbrace{L(f_\delta^* - g_\delta^*)}_{=-\omega(\delta)} \\ &= \omega(\delta) - \omega'(\delta)\delta. \end{aligned}$$

Then T_δ minimizes $\text{var}(T)$ among affine statistics $T = a + \langle k, Y\rangle$ subject to the constraint

$$\text{for all } f \in \mathcal{F}, g \in \mathcal{G},\ \text{bias}_f(T) - \text{bias}_g(T) \le B.$$

Any other statistic solving this minimization problem takes the form $T_\delta + a$ for some a.

pf.: The result follows by showing that any other statistic would lead to a test that is more powerful than the NP test of f^* vs g^*, which would require that this statistic lead to the same test as T_δ. Formally, let $k_\delta^* = \frac{\omega(\delta)}{\delta} K(g_\delta^* - f_\delta^*)$ and let $T = a + \langle k, Y\rangle$ be a statistic that solves this optimization problem. The lemma is equivalent to showing $k = k_\delta^*$. By optimality of T, $\text{bias}_{f^*}(T) - \text{bias}_{g^*}(T) \le B$ and $\text{var}(T) \le \text{var}(T_\delta)$. This leads to a level α test of g^* based on T: reject when $T > Lf^* + \text{bias}_{f^*}(T) + z_{1-\alpha}\text{std}(T)$, which would have power

$$\Phi\left(\frac{L(g^* - f^*) + \overbrace{\text{bias}_{g^*}(T) - \text{bias}_{f^*}(T)}^{\ge -B}}{\text{std}(T)} - z_{1-\alpha}\right) \tag{*}$$

since $T \overset{g^*}{\sim} N(Lg^* + \text{bias}_{g^*}(T), \text{var}(T))$. The NP test of f^* vs g^* rejects for large values of T_δ and has power

$$\Phi\left(\frac{L(g^* - f^*) + \overbrace{\text{bias}_{g^*}(T_\delta) - \text{bias}_{f^*}(T_\delta)}^{=-B}}{\text{std}(T_\delta)} - z_{1-\alpha}\right). \tag{**}$$

The Neyman-Pearson lemma implies (**) $\geq$ (*), which means that

$$\text{var}(T) = \text{var}(T_\delta) \quad \text{and} \quad \text{bias}_{f^*}(T) - \text{bias}_{g^*}(T) = B = \text{bias}_{f^*}(T_\delta) - \text{bias}_{g^*}(T_\delta)$$

(note that the numerator in the display is positive, since $L(g^* - f^*) = \omega(\delta) \geq B$). These equalities can be rewritten

$$\|k\|^2 = \|k_\delta^*\|^2 \quad \text{and} \quad \langle k, k_\delta^* \rangle = \langle k_\delta^*, k_\delta^* \rangle$$

where the latter follows since $\text{bias}_{f^*}(T) - \text{bias}_{g^*}(T) = \text{bias}_{f^*}(T_\delta) - \text{bias}_{g^*}(T_\delta)$ iff. $\langle k, K(f^* - g^*) \rangle = \langle k_\delta^*, K(f^* - g^*) \rangle$ and $K(f^* - g^*)$ is proportional to k_δ^*. This implies $\|k - k_\delta^*\|^2 = \|k_\delta^*\|^2 + \|k\|^2 - 2\langle k, k_\delta^* \rangle = 0$, which gives the result.

- Note:
  - The above lemma still holds if ε is non-normal (the proof is the same, with statements of the form "the Neyman-Pearson test is..." changed to "the Neyman-Pearson test with ε normal would be...").
  - The above lemma can be strengthened to show that T_δ is minimum variance among all (not just affine) estimators satisfying the bias constraints. (This requires normality. See Low 1995.)

- The lemma characterizes tradeoffs between worst-case upward bias over $\mathcal{F}$, worst case downward bias over $\mathcal{G}$, and variance.

- Let $\overline{\text{bias}}_{\mathcal{F}}(T) = \sup_{f \in \mathcal{F}} \text{bias}_f(T)$ and $\underline{\text{bias}}_{\mathcal{G}}(T) = \inf_{g \in \mathcal{G}} \text{bias}_g(T)$. If an estimator T satisfies $\overline{\text{bias}}_{\mathcal{F}}(T) \leq \bar{b}$ and $\underline{\text{bias}}_{\mathcal{G}}(T) \geq \bar{b} - B$, then it cannot have variance below $\omega'(\delta)^2 = \text{var}(T_\delta)$. An estimator achieving this variance is given by $a + T_\delta$ with a calibrated so that $\overline{\text{bias}}_{\mathcal{F}}(a + T_\delta) = \bar{b}$.

- Let us calibrate a so that

$$\text{bias}_{f^*}(a + T_\delta) = \overline{\text{bias}}_{\mathcal{F}}(a + T_\delta) = -\underline{\text{bias}}_{\mathcal{G}}(a + T_\delta) = -\text{bias}_{g^*}(a + T_\delta) = B/2$$

Define $f_{M,\delta}^* = (g_\delta^* + f_\delta^*)/2$. Then the above display implies

$$0 = \text{bias}_{f_M^*}(a + T_\delta) = a + \frac{\omega'(\delta)}{\delta} \langle K(f_\delta^* - g_\delta^*), K f_{M,\delta}^* \rangle - L f_{M,\delta}^*$$

which gives the estimator

$$\hat{L}_\delta \equiv a + T_\delta = Lf^*_{M,\delta} + \frac{\omega'(\delta)}{\delta}\langle K(f^*_\delta - g^*_\delta), Y - Kf^*_{M,\delta}\rangle.$$

- Summary of properties of $\hat{L}_\delta$:
  - $\text{bias}_{f^*}(\hat{L}_\delta) = \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta) = -\underline{\text{bias}}_{\mathcal{G}}(\hat{L}_\delta) = -\text{bias}_{g^*}(\hat{L}_\delta) = \frac{1}{2}(\omega(\delta) - \omega'(\delta)\delta)$
  - $\text{var}(\hat{L}_\delta) = \omega'(\delta)^2$
  - $\hat{L}_\delta$ is the unique choice of T that minimizes $\text{var}(T)$ among affine estimates with $\overline{\text{bias}}_{\mathcal{F}}(T) \leq \frac{1}{2}(\omega(\delta) - \omega'(\delta)\delta)$ and $\underline{\text{bias}}_{\mathcal{G}}(T) \geq -\frac{1}{2}(\omega(\delta) - \omega'(\delta)\delta)$.
  - The NP test of f^*_δ vs g^*_δ (which, by the convex testing lemma, is also the minimax test of $H_0 : Lf \leq Lf^*_\delta$, $f \in \mathcal{F}$ vs $H_1 : Lf \geq Lf^*_\delta + \omega(\delta)$, $f \in \mathcal{G}$) rejects for large values of $\hat{L}_\delta$.

3.1 One-sided adaptation

- Consider one-sided CIs of the form $\mathcal{C} = [\hat{c}, \infty)$.
- Among CIs with a given coverage level, we want smaller values of excess length $Lf - \hat{c}$ (or, perhaps, $(Lf - \hat{c})_+$).
- Let $q_{\beta,f}(T)$ denote the β quantile of T under f.
- Let

$$q_\beta(\hat{c}, \mathcal{G}) = \sup_{g \in \mathcal{G}} q_{\beta,g}(Lg - \hat{c})$$

  denote the worst-case β quantile of excess length over a set $\mathcal{G}$.

- We will show that the adaptive CI problem

$$\min_{\hat{c}} q_\beta(\hat{c}, \mathcal{G}) \text{ s.t. } \inf_{f \in \mathcal{F}} P_f(Lf \in [\hat{c}, \infty)) \qquad (*)$$

  is solved by a CI based on $\hat{L}_\delta$ for δ appropriately calibrated.

- Given α and β, let $\delta = z_\beta + z_{1-\alpha}$ so that the NP test of f^*_δ vs g^*_δ (which is is based on $\hat{L}_\delta$) has power $\beta = \Phi(\delta - z_{1-\alpha})$.

- This leads to $1-\alpha$ CI $[\hat{c}_{\delta,\alpha}, \infty)$ where

$$\begin{aligned}\hat{c}_{\delta,\alpha} &= \hat{L}_\delta - \overline{\text{bias}}_{\mathcal{F}}(\hat{L}_\delta) - z_{1-\alpha}\text{std}(\hat{L}_\delta) \\ &= \hat{L}_\delta - \frac{1}{2}\left(\omega(\delta) - \omega'(\delta)\delta\right) - z_{1-\alpha}\omega'(\delta)\end{aligned}$$

  The worst-case β quantile excess length is

$$q_\beta(\hat{c}_{\delta,\alpha}, \mathcal{G}) = \sup_{g\in\mathcal{G}} q_{g,\beta}(Lg - \hat{c}_{\alpha,\delta}) = q_{g^*_\delta,\beta}(Lg^*_\delta - \hat{c}_{\alpha,\delta}) = \omega(\delta)$$

  (the last step can be verified directly or by noting that the NP test of f^*_δ vs g^*_δ has power β and rejects when $Lf^*_\delta \leq \hat{c}_{\alpha,\delta} \iff Lg^*_\delta - \hat{c}_{\alpha,\delta} \leq \omega(\delta)$).

- thm.: Suppose that $\mathcal{G} \subseteq \mathcal{F}$ and that the modulus is achieved at $\delta = z_\beta + z_{1-\alpha}$ with $\|K(f^*_\delta - g^*_\delta)\| = \delta$. Then $\hat{c}_{\delta,\alpha}$ solves (*) and achieves $q_\beta(\hat{c}_{\delta,\alpha}, \mathcal{G}) = \omega(\delta)$. Coverage is minimized at f^*_δ and all quantiles of excess length are maximized at g^*_δ.

  pf.: The results all follow from derivations above except for the claim that $\hat{c}_{\alpha,\delta}$ solves (*). This follows by showing that, if another CI achieved strictly shorter βth quantile excess length at g^*_δ, it could be used to achieve a test of $H_0 : (1-\lambda)f^*_\delta + \lambda g^*_\delta$ vs $H_1 : g^*_\delta$ with power strictly greater than the NP test for some $\lambda \in (0,1)$, which would contradict the Neyman-Pearson lemma.

3.2 Centrosymmetry and translation invariance

- The modulus simplifies when $\mathcal{F}$ and $\mathcal{G}$ satisfy certain properties.
- def.: $\mathcal{F}$ is centrosymmetric (CS) if $f \in \mathcal{F} \implies -f \in \mathcal{F}$.
- def.: $\mathcal{F}$ is translation invariant (TI) if there exists ι such that $L\iota = 1$ and $f + c\iota \in \mathcal{F}$ for all $c \in \mathbb{R}$, $f \in \mathcal{F}$.
- Under translation invariance, we have $\omega'(\delta) = \frac{\delta}{\langle K(g^*_\delta - f^*_\delta), K\iota\rangle}$ (see lemma below) so that

$$\hat{L}_\delta = Lf^*_{M,\delta} + \frac{\langle K(g^*_\delta - f^*_\delta), Y - Kf^*_{M,\delta}\rangle}{\langle K(g^*_\delta - f^*_\delta), K\iota\rangle}.$$

- Under centrosymmetry, if f^*_δ and g^*_δ solve the single class modulus (where $\mathcal{G} = \mathcal{F}$), then $(f^*_\delta - g^*_\delta)/2$ and $-(f^*_\delta - g^*_\delta)/2$ also solve the single class modulus. Thus, we can restrict attention to solutions with $g^*_\delta = -f^*_\delta$, which gives

$$\omega(\delta;\mathcal{F},\mathcal{F}) = \sup\left\{2Lf | \|Kf\| \le \delta/2, f \in \mathcal{F}\right\}$$

and, using the fact that $f^*_{M,\delta} = 0$ if $g^*_\delta = -f^*_\delta$,

$$\hat{L}_\delta = \frac{2\omega'(\delta)}{\delta}\langle Kg^*_\delta, Y\rangle \overset{TI}{=} \frac{\langle Kg^*_\delta, Y\rangle}{\langle Kg^*_\delta, K\iota\rangle}.$$

- In the fixed design regression model, if TI holds with $\iota(x) = 1$, the last expression (where TI and CS both hold) is a Nadaraya-Watson estimator: $\frac{\sum_{i=1}^n g^*_\delta(x_i)Y_i}{\sum_{i=1}^n g^*_\delta(x_i)}$.

- Lemma: Suppose that $\|K(g^*_\delta - f^*_\delta)\| = \delta$ (the constraint is binding in the modulus problem) and $f^*_\delta + c\iota \in \mathcal{F}$ for all c in a neighborhood of zero, where $L\iota = 1$. Then $\partial\omega(\delta) = \left\{\frac{\delta}{\langle K(g^* - f^*), K\iota\rangle}\right\}$.

  pf.: Let $d \in \partial\omega(\delta)$ and let $f_c = f^* - c\iota$. Let η be small enough so that $f_c \in \mathcal{F}$ for $|c| < \eta$. Then, for $|c| \le \eta$,

$$\begin{aligned}
&L(g^* - f^*) + d\left[\|K(g^* - f_c)\| - \delta\right] \\
&\ge \omega(\|K(g^* - f_c)\|) && \text{(def. of superdifferential)} \\
&\ge L(g^* - f_c) && \text{def. of modulus} \\
&= L(g^* - f^*) + c\underbrace{L\iota.}_{=1}
\end{aligned}$$

Since the two sides of the above display are equal at $c = 0$ and the left hand side is greather than or equal to the right hand side, the derivatives at 0 are equal, which gives

$$1 = d \cdot \frac{d}{dc}\|K(g^* - f_c)\|\bigg|_{c=0} = d \cdot \frac{\frac{d}{dc}\|K(g^* - f_c)\|^2\Big|_{c=0}}{2\delta} = d \cdot \frac{\langle K(g^* - f^*), K\iota\rangle}{\delta}$$

so that $d = \frac{\delta}{\langle K(g^* - f^*), K\iota\rangle}$ as claimed.

3.3 Solving the modulus problem/examples

- The modulus problem is a convex optimization problem in $\mathcal{F}$, which is often infinite dimensional. However, all that really matters is Kf, which is often in a finite dimensional space (e.g. $\mathbb{R}^n$ in the fixed design regression model). If we can phrase the constraints on $\mathcal{F}$ as constraints on Kf, then we can at least reduce this to an optimization problem over a finite dimensional space. Depending on how we phrase the constraints, the problem may still be convex, making it computationally feasible even for n large.

- Often, the structure of the problem gives a simpler solution.

- The problem is also related to the problem of optimal recovery from approximation theory/computer science, so results from that literature can be helpful (Donoho, 1994, gives some references to this literature).

- Let us cover some examples, focusing on the minimax ($\mathcal{G} = \mathcal{F}$) (see Donoho 1994, Lepski and Tsybakov 2000 and references therein for more on these and other examples).

- Smoothness classes of functions $f : \mathcal{X} \to \mathbb{R}$: Let $\gamma \in [0, \infty)$, $p \in \mathbb{N}$, $\mathcal{X} \subseteq \mathbb{R}$.

  - Hölder class:

    $$\mathcal{F}_H(\gamma, C) = \left\{ f \,\middle|\, \left| f^{(\ell)}(x) - f^{(\ell)}(x') \right| \le C|x - x'|^{\gamma - \ell} \text{ all } x, x' \in \mathcal{X} \right\}$$

    where ℓ is the maximum integer strictly less than γ.

  - Sobolev class:

    $$\mathcal{F}_S(p, C) = \left\{ f \,\middle|\, \int_{\mathcal{X}} (f^{(p)}(x))^2 \, dx \le C^2 \right\}$$

    (Note: sometimes Sobolev classes are defined in terms of $\int (f^{(p)})^q$ for q possibly not equal to 2.)

  - Taylor class at x_0:

    $$\mathcal{F}_T(p, C) = \left\{ f \,\middle|\, \left| f(x) - \sum_{j=0}^{p-1} f^{(j)}(x_0)(x - x_0)^j / j! \right| \le C|x - x_0|^p \text{ all } x \in \mathcal{X} \right\}$$

- Note: can take ι to be a $p-1$th (for Taylor and Sobolev) or ℓth order (for Hölder) polynomial for translation invariance

- Transformation K: see examples from Section 2.3 (fixed design regression, Gaussian white noise, etc.)

- Functional L: recall examples from Section 1 (all of them are linear)
  - $Lf = f(x_0)$
  - $Lf = f^{(r)}(x_0)$
  - RD: $Lf = \lim_{x \downarrow x_0} f(x) - \lim_{x \uparrow x_0} f(x)$
  - ATE under unconfoundedness: $Lf = \frac{1}{n}\sum_{i=1}^n [f(w_i, 1) - f(w_i, 0)]$ where $x_i = (w_i, d_i)$, d_i an indicator for "treatment"

3.3.1 Example: $Lf = f(0)$, $\mathcal{F} = \mathcal{G} = \mathcal{F}_H(\gamma, C)$, $\gamma \le 1$

- Consider fixed design regression ($Kf = (f(x_1)/\sigma(x_1), \ldots, f(x_n)/\sigma(x_n))$).

- Easier to use inverse modulus (dual of this problem). Using formula under CS, this is

$$\min 2\sqrt{\sum_{i=1}^n \frac{f(x_i)^2}{\sigma^2(x_i)}} \text{ s.t. } 2f(0) \ge b \text{ and for all } x, x' \in \mathbb{R},\ |f(x) - f(x')| \le C|x - x'|.$$

- Subject to the constraint, we can make $|f(x)|$ as small as possible simultaneously for all x by setting

$$f(x) = \max\{b/2 - C|x|^\gamma, 0\}.$$

  so $f^*_{\omega^{-1}(b)}(x)$ is given by the function in the above display.

- Can take $\iota(x) = 1$, which gives

$$\hat{L}_{\omega^{-1}(b)} = \frac{\sum_{i=1}^n y_i \max\{b/2 - C|x_i|^\gamma, 0\}/\sigma^2(x_i)}{\sum_{i=1}^n \max\{b/2 - C|x_i|^\gamma, 0\}/\sigma^2(x_i)} = \frac{\sum_{i=1}^n y_i \max\{1 - |x_i/h|^\gamma, 0\}/\sigma^2(x_i)}{\sum_{i=1}^n \max\{1 - |x_i/h|^\gamma, 0\}/\sigma^2(x_i)}$$

  where $h = (2C/b)^{1/\gamma}$.

- To get f^*_δ, $\hat{L}_\delta$, calibrate b so that $2\sqrt{\sum_{i=1}^n \max\{b/2 - C|x_i|^\gamma, 0\}^2/\sigma^2(x_i)} = \delta$.

3.3.2 Example: Other functionals with $\mathcal{F} = \mathcal{G} = \mathcal{F}_H(1, C)$

- Sometimes the solution is difficult to characterize, but can be reduced to a finite dimensional convex progamming problem.
- Generalize the Lipschitz class ($\mathcal{F}_H(1, C)$) to arbitrary metric space $\mathcal{X}$ (usually $\mathbb{R}^n$) with distance $d_{\mathcal{X}}(x, x')$:

$$\mathcal{F}_{Lip}(C) = \{f : \mathcal{X} \to \mathbb{R} || f(x) - f(x')| \leq C d_{\mathcal{X}}(x, x')\}.$$

- Let L have the form

$$Lf = \sum_{j=1}^{m} w(\tilde{x}_j) f(\tilde{x}_j)$$

  where $w(\cdot)$ is a known weighting function and $\{\tilde{x}_j\}_{j=1}^m$ are known.
- Interpretation: in the case where $w(\tilde{x}) = 1/m$, we are estimating the average of the expectation of the outcome y for individuals $j = 1, \ldots, m$ described by covariates $\tilde{x}_j$, using data on individuals $i = 1, \ldots, n$ described by covariates x_i. With other choices of $w(\cdot)$, we can get differences of these averages (i.e. average treatment effects), or weighted versions of them.
  - Sample ATE: $Lf = \frac{1}{n} \sum_{i=1}^{n} [f(v_i, 1) - f(v_i, 0)]$ where $x_i = (v_i, d_i)$, we can set $w(v, d) = (2d - 1)/n$, $m = 2n$ and $\tilde{x}_j = (v_j, 1)$, $\tilde{x}_{n+j} = (v_j, 0)$ for $j = 1, \ldots, n$.
- Modulus problem (using centrosymmetry):

$$\max_{f:\mathcal{X}\to\mathbb{R}} 2 \sum_{j=1}^{m} w(\tilde{x}_j) f(\tilde{x}_j) \quad \text{s.t.} \quad \sum_{i=1}^{n} \frac{f(x_i)^2}{\sigma^2(x_i)} \leq \frac{\delta^2}{4}, \text{ and}$$
$$\text{for all } x, x' \in \mathcal{X},\ |f(x) - f(x')| \leq C d_{\mathcal{X}}(x, x').$$

- Let $\widetilde{\mathcal{X}} = \widetilde{\mathcal{X}}_{n,m} = \{x_1, \ldots, x_n\} \cup \{\tilde{x}_1, \ldots, \tilde{x}_m\}$ and let $f^* : \widetilde{\mathcal{X}}_{n,m} \to \mathbb{R}$ solve the same problem with $\mathcal{X}$ replaced by $\widetilde{\mathcal{X}}_{n,m}$:

$$\max_{f:\widetilde{\mathcal{X}}_{n,m}\to\mathbb{R}} 2\sum_{j=1}^{m} w(\tilde{x}_j)f(\tilde{x}_j) \quad \text{s.t.} \quad \sum_{i=1}^{n}\frac{f(x_i)^2}{\sigma^2(x_i)} \leq \frac{\delta^2}{4}, \text{ and}$$
$$\text{for all } x, x' \in \widetilde{\mathcal{X}}_{n,m},\ |f(x) - f(x')| \leq Cd_{\mathcal{X}}(x, x').$$

- By Theorem 4 of Beliakov (2006), it is possible to extend f^* to a function on $\mathcal{X}$ that satisfies the Lipschitz constraint for all $x, x' \in \mathcal{X}$ (not just $\widetilde{\mathcal{X}}_{n,m}$). Since this function achieves the modulus with the relaxed constraints (Lipschitz only on $\widetilde{\mathcal{X}}_{n,m}$) and satisfies the full constraints (Lipschitz on $\mathcal{X}$), it must solve the modulus problem.
- Note that we do not have to find this function explicitly, since $\omega(\delta)$ and $\hat{L}_\delta$ depend only on $f^*_\delta(x)$ for $x \in \widetilde{\mathcal{X}}_{n,m}$.
- This reduces the modulus problem to maximizing a linear function in $\#\widetilde{\mathcal{X}}_{n,m} \leq n + m$ dimensional space subject to a convex quadradic constraint and $\#\widetilde{\mathcal{X}}_{n,m}\cdot\left(\#\widetilde{\mathcal{X}}_{n,m} - 1\right)/2$ linear inequality constraints. It can be solved using convex optimization packages such as CVX for Matlab.

3.3.3 Example: approximately linear models (Sacks and Ylvisaker, 1978)

- Observe $\{(x_i, y_i)\}_{i=1}^n$ where

$$y_i = x_i'\gamma + c_i + u_i$$

where $u = (u_1, \ldots, u_n)' \sim N(0, \Sigma)$, $\gamma \in \mathbb{R}^k$ and $|c_i| \leq r_i$. Parameter space $\mathcal{F}$ is given by $\{(\gamma', c')' | \gamma \in \mathbb{R}^k, c \in [-r_1, r_1] \times \cdots \times [-r_n, r_n]\}$. Let $Y = \Sigma^{-1/2}(y_1, \ldots, y_n)'$, $K(\gamma', c')' = \Sigma^{-1/2}(X\gamma + c)$ where $\underset{n\times k}{X} = (x_1, \ldots, x_n)'$, and let $L(\gamma', c')' = \ell'\gamma$ for some $\ell \in \mathbb{R}^k$.

- One-class modulus (using centrosymmetry)

$$\max_{\gamma, c} 2\ell'\gamma \quad \text{s.t.} \quad (X\gamma + c)'\Sigma^{-1}(X\gamma + c) \leq \delta^2/4, \text{ and } |c_i| \leq r_i \text{ all } i = 1, \ldots, n.$$

This is a finite dimensional convex programming problem. The class is translation invariant with $\iota = (\ell'/\|\ell\|^2, 0)'$ (i.e. $c = 0$), which gives

$$\hat{L}_\delta = \frac{(X\gamma^*_\delta + c^*_\delta)'\Sigma^{-1}(y_1, \ldots, y_n)'}{(X\gamma^*_\delta + c^*_\delta)'\Sigma^{-1}X\ell/\|\ell\|^2}$$

- When $\Sigma = diag(\sigma^2(x_1), \ldots, \sigma^2(x_n))$ (independent observations) we get

$$\max_{\gamma,c} 2\ell'\gamma \quad \text{s.t.} \quad \sum_{i=1}^{n}(x_i'\gamma + c_i)^2/\sigma^2(x_i) \leq \delta^2/4, \text{ and } |c_i| \leq r_i \text{ all } i = 1, \ldots, n.$$

- We can minimize each $|x_i'\gamma + c_i|$ individually by setting $c_i = -x_i'\gamma$ if $x_i'\gamma \in [-r_i, r_i]$ and setting $c_i = -r_i \cdot \text{sign}(x_i'\gamma)$ if $x_i'\gamma \notin [-r_i, r_i]$. Then $x_i'\gamma + c_i = (x_i'\gamma - r_i)_+ - (x_i'\gamma + r_i)_-$ where $(t)_+ = \max\{t, 0\}$ and $(t)_- = \max\{-t, 0\}$. Thus, we can optimize over γ only. Estimator is

$$\hat{L}_\delta = \frac{\sum_{i=1}^n [(x_i'\gamma^*_\delta - r_i)_+ - (x_i'\gamma^*_\delta + r_i)_-] y_i/\sigma^2(x_i)}{\sum_{i=1}^n [(x_i'\gamma^*_\delta - r_i)_+ - (x_i'\gamma^*_\delta + r_i)_-] x_i'\ell/(\|\ell\|^2\sigma^2(x_i))}$$

- The Taylor class $\mathcal{F}_T(p, C)$ falls into this framework. Let $x_0 = 0$ for simplicity. Then $f(x_i) = \sum_{j=0}^p f^{(j)}(0)x_i^j/j! + c_i$ where $|c_i| \leq C|x_i|^p$. Thus, we get the approximately linear model with $r_i = C|x_i|^p$, $(1, x_i, x_i^2, \ldots, x_i^{(p-1)})'$ playing the role of x_i and with $(f(0), f'(0), f''(0)/2, \ldots, f^{(p-1)}(0)/(p-1)!)'$ playing the role of γ.

- Historical note: Estimation of a conditional mean at the boundary was one of the original motivations in Sacks and Ylvisaker (1978), who obtain finite sample results in the fixed design regression model. Cheng, Fan, and Marron (1997) use their results to characterize asymptotic relative efficiencies of local linear estimators for boundary estimation. While the latter paper is often cited in the RD literature to justify the use of local linear estimators for this problem, the finite sample approach of Sacks and Ylvisaker (1978) has, to my knowledge, only been applied to this problem recently by Armstrong and Kolesár (2015). Although they do not mention it in their paper, Sacks and Ylvisaker were motivated by RD as well and wrote their paper after discussing the problem with Donald Campbell, one of the coauthors of the seminal paper (Thistlethwaite and Campbell, 1960) that introduced the method. See historical accounts in Cook (2008) and Sacks and Ylvisaker (2012).

3.4 Limits to adaptation under centrosymmetry

- Let $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}}$ denote the solution to

  $$\min_{\hat{c}} q_\beta(\hat{c},\mathcal{G}) \text{ s.t. } \inf_{f\in\mathcal{F}} P_f(Lf \in [\hat{c},\infty)) \geq 1-\alpha$$

  derived in Section 3.1 (denoted $\hat{c}_{\alpha,\delta}$ in that section), and let $f^*_{\delta,\mathcal{F},\mathcal{G}}$ and $g^*_{\delta,\mathcal{F},\mathcal{G}}$ denote least favorable functions.

- Suppose that $\mathcal{F}$ is centrosymmetric. For the modulus problem for $\omega(2\delta,\mathcal{F},\mathcal{F})$, we can take $g^*_{2\delta,\mathcal{F},\mathcal{F}} = -f^*_{2\delta,\mathcal{F},\mathcal{F}}$ where $f^*_{2\delta,\mathcal{F},\mathcal{F}}$ solves

  $$\omega(2\delta;\mathcal{F},\mathcal{F}) = \sup\{-2Lf | \|Kf\| \leq \delta,\, f\in\mathcal{F}\}.$$

  Consider adapting to the class $\{0\}$ (the zero function). The least favorable function $f^*_{\delta,\mathcal{F},\{0\}}$ (with δ instead of 2δ) solves

  $$\omega(\delta;\mathcal{F},\{0\}) = \sup\{-Lf | \|Kf\| \leq \delta,\, f\in\mathcal{F}\}.$$

  Thus,

  $$\omega(\delta;\mathcal{F},\{0\}) = \frac{1}{2}\omega(2\delta;\mathcal{F},\mathcal{F}) \quad \text{and} \quad f^*_{\delta,\mathcal{F},\{0\}} = f^*_{2\delta,\mathcal{F},\mathcal{F}} = \frac{f^*_{2\delta,\mathcal{F},\mathcal{F}} - g^*_{2\delta,\mathcal{F},\mathcal{F}}}{2}.$$

  The latter result means that $\hat{c}_{\alpha,\delta,\mathcal{F},\{0\}} = \hat{c}_{2\delta,\mathcal{F},\mathcal{F}}$ (assuming the same element in $\partial\omega(\delta)$ is used) since f^* is the same for both modulus problems and $g^* - f^*$ is the same up to scale.

- Thus, minimax CI for $\beta = \Phi(2\delta - z_{1-\alpha})$ is identical to CI "directed at 0" for $\beta = \Phi(\delta - z_{1-\alpha})$.

- Consider performance of $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{F}}$ (minimax CI for quantile $\beta = \Phi(\delta - z_{1-\alpha})$) at $f = 0$. Note that

  $$\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{F}} = \hat{L}_{\delta,\mathcal{F},\mathcal{F}} - \frac{1}{2}\left(\omega(\delta;\mathcal{F},\mathcal{F}) - \omega'(\delta;\mathcal{F},\mathcal{F})\delta\right) - z_{1-\alpha}\omega'(\delta,\mathcal{F},\mathcal{F}).$$

Since $f^*_{M,\delta,\mathcal{F},\mathcal{F}} = 0$, the bias at $f = 0$ is 0, so that

$$Lf - \hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{F}} \overset{f=0}{\sim} N\left(\frac{1}{2}\left(\omega(\delta;\mathcal{F},\mathcal{F}) - \omega'(\delta;\mathcal{F},\mathcal{F})\delta\right) + z_{1-\alpha}\omega'(\delta;\mathcal{F},\mathcal{F}), [\omega'(\delta;\mathcal{F},\mathcal{F})]^2\right)$$

which gives, for $\beta = \Phi(\delta - z_{1-\alpha})$,

$$\begin{aligned} q_\beta(\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{F}}, \{0\}) &= \underbrace{(z_\beta + z_{1-\alpha})}_{=\delta}\omega'(\delta;\mathcal{F},\mathcal{F}) + \frac{1}{2}\left(\omega(\delta;\mathcal{F},\mathcal{F}) - \omega'(\delta;\mathcal{F},\mathcal{F})\delta\right) \\ &= \frac{1}{2}\left(\omega(\delta;\mathcal{F},\mathcal{F}) + \omega'(\delta;\mathcal{F},\mathcal{F})\delta\right). \end{aligned}$$

The CI $[\hat{c}_{\alpha,\delta,\mathcal{F},\{0\}})$ optimizes $q_\beta(\hat{c}, \{0\})$ and achieves $q_\beta(\hat{c}_{\alpha,\delta,\mathcal{F},\{0\}}, \{0\}) = \omega(\delta;\mathcal{F},\{0\}) = \frac{1}{2}\omega(2\delta;\mathcal{F},\mathcal{F})$. Thus, "directed at $\{0\}$" CI improves upon minimax CI by a factor of

$$\frac{\omega(2\delta;\mathcal{F},\mathcal{F})}{\omega(\delta;\mathcal{F},\mathcal{F}) + \omega'(\delta;\mathcal{F},\mathcal{F})\delta}$$

under $f = 0$.

- Graphical interpretation: denominator is the Taylor approximation to $\omega(2\delta)$ (the numerator) expanding at δ.

- Note that the bound is $\geq 1/2$ by concavity and nonnegativity of ω.

- Now consider any class $\mathcal{G}$ such that

$$f - g^*_{\delta,\mathcal{F},\mathcal{G}} \in \mathcal{F} \text{ all } f \in \mathcal{F}. \qquad (*)$$

  Then $\omega(\delta;\mathcal{F},\mathcal{G}) = \omega(\delta;\mathcal{F},\{g^*_{\delta,\mathcal{F},\mathcal{G}}\}) = \omega(\delta;\mathcal{F},\{0\})$ and the pair $f^*_{\delta,\mathcal{F},\mathcal{G}} - g^*_{\delta,\mathcal{F},\mathcal{G}}$, 0 solve $\omega(\delta;\mathcal{F},\{0\})$ (the last step follows since, under (*), $f - g^*_{\delta,\mathcal{F},\mathcal{G}} \in \mathcal{F}$ iff. $f \in \mathcal{F}$). Thus, $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}} = \hat{c}_{\alpha,\delta,\mathcal{F},\{0\}}$, and the arguments apply to CIs that are "directed" at $\mathcal{G}$ as well. This gives ...

- thm.: Let $\delta = z_\beta + z_{1-\alpha}$. Suppose that $\mathcal{F}$ and $\mathcal{G}$ satisfy (*) and $\mathcal{G} \subseteq \mathcal{F}$ and $\|K(f^*_{\delta,\mathcal{F},\mathcal{G}} - g^*_{\delta,\mathcal{F},\mathcal{G}})\| = \delta$. Then, setting $\tilde{\beta} = \Phi((z_\beta - z_{1-\alpha})/2)$ (so that $\delta/2 = z_{\tilde{\beta}} + z_{1-\alpha}$), the minimax β quantile CI $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{F}}$ optimizes $q_{\tilde{\beta}}(\hat{c},\mathcal{G})$. The efficiency of the minimax β quantile CI

$\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{F}}$ for worst case β quantile over $\mathcal{G}$ is

$$\frac{q_\beta(\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}},\mathcal{G})}{q_\beta(\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{F}},\mathcal{G})} = \frac{\omega(2\delta;\mathcal{F},\mathcal{F})}{\omega(\delta;\mathcal{F},\mathcal{F}) + \delta\omega'(\delta;\mathcal{F},\mathcal{F})}.$$

- Summary of lack-of-adaptation for centrosymmetric $\mathcal{F}$
  - Minimax CI at quantile $\beta = \Phi(\delta - z_{1-\alpha})$ has worst-case β quantile excess length $\omega(\delta;\mathcal{F},\mathcal{G})$.
  - The same CI optimizes worst-case $\tilde{\beta} = \Phi((z_\beta - z_{1-\alpha})/2)$ quantile excess length over $\mathcal{G}$ satisfying (*).
  - For βth quantile excess length over $\mathcal{G}$, $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{F}}$ optimizes the "wrong quantile," which gives the worst-case excess length over $\mathcal{G}$ as $\frac{1}{2}(\omega(\delta;\mathcal{F},\mathcal{F}) + \omega'(\delta;\mathcal{F},\mathcal{F})\delta)$
  - If we optimize βth quantile excess length instead, we use $\hat{c}_{\alpha,\delta,\mathcal{F},\mathcal{G}} = \hat{c}_{\alpha,2\delta,\mathcal{F},\mathcal{F}}$, which gives worst-case β quantile excess length over $\mathcal{G}$ as $\omega(2\delta;\mathcal{F},\mathcal{G}) = \frac{1}{2}\omega(2\delta;\mathcal{F},\mathcal{F})$.
- Graphically, we can draw the tangent line to $\omega(\cdot;\mathcal{F},\mathcal{F})$ at δ. Then, take the point on this line where the x-axis is 2δ and draw a line from there to the origin. The point on this line where the x-axis is δ gives $\frac{1}{2}(\omega(\delta;\mathcal{F},\mathcal{F}) + \omega'(\delta;\mathcal{F},\mathcal{F})\delta)$. Now, draw a line from the origin to $(2\delta,\omega(2\delta;\mathcal{F},\mathcal{F}))$ and take the point on this line where the x-axis is δ. This gives $\frac{1}{2}\omega(2\delta;\mathcal{F},\mathcal{F})$.

3.4.1 Example: monotonicity restrictions

- The class $\mathcal{F} = \mathcal{F}_H(\gamma, C)$ is centrosymmetric, so the lack of adaptivity results apply. In particular, if we optimize the worst-case $\beta = \Phi(\delta - z_{1-\alpha})$ quantile over $\mathcal{G} = \mathcal{F}_H(\gamma_2, C)$ subject to coverage over $\mathcal{F} = \mathcal{F}_H(\gamma_1, C)$, we get a CI with worst-case β quantile

$$\omega(\delta;\mathcal{F}_H(\gamma_1,C),\mathcal{F}_H(\gamma_2,C)) \geq \omega(\delta;\mathcal{F}_H(\gamma_1,C),\{0\}) = \frac{1}{2}\omega(2\delta;\mathcal{F}_H(\gamma_1,C),\mathcal{F}_H(\gamma_1,C)).$$

- However, the class $\mathcal{F} = \mathcal{F}_H(\gamma, C) \cap \{f \text{ nonincreasing}\}$ is not centrosymmetric.
- Consider fixed design regression with $Lf = f(0)$, $\mathcal{F} = \mathcal{F}_H(\gamma_1, C) \cap \{f \text{ nonincreasing}\}$ and $\mathcal{G} = \mathcal{F}_H(\gamma_2, C) \cap \{f \text{ nonincreasing}\}$ with $\gamma_\ell \leq \gamma_u \leq 1$.

- Inverse modulus problem

$$\min \sum_{i=1}^{n} (g(x_i) - f(x_i))^2/\sigma^2(x_i) \text{ s.t. } g(0) - f(0) \geq b, f \in \mathcal{F}_H(\gamma_1, C),\ g \in \mathcal{F}_H(\gamma_2, C),$$
$$f \text{ and } g \text{ nonincreasing.}$$

- By translation invariance, we can fix $f(0)$ at an arbitrary point (say $f(0) = 0$). Least favorable functions are

$$f^*(x) = \begin{cases} b & x < 0 \text{ and } C|x|^{\gamma_1} \geq b \\ C|x|^{\gamma_1} & x < 0 \text{ and } C|x|^{\gamma_1} \leq b \\ 0 & x \geq 0 \end{cases}$$

  and

$$g^*(x) = \begin{cases} b & x \leq 0 \\ b - C|x|^{\gamma_2} & x > 0 \text{ and } C|x|^{\gamma_2} \leq b \\ 0 & x \geq 0 \text{ and } C|x|^{\gamma_2} > b \end{cases}$$

- Using formula under TI, we get

$$\hat{L}_{\omega^{-1}(b),\mathcal{F},\mathcal{G}} = \frac{\sum_{i=1}^n y_i k(x_i)/\sigma^2(x_i)}{\sum_{i=1}^n k(x_i)/\sigma^2(x_i)} \quad \text{where} \quad k(x) = \begin{cases} \max\{b - C|x|^{\gamma_1}, 0\} & x \leq 0 \\ \max\{b - C|x|^{\gamma_2}, 0\} & x \geq 0 \end{cases}$$

- Note that the number of observations to the left of zero depends only on on γ_1 (null smoothness), and the number of observations to the right depends only on γ_2 (alternative smoothness). As $b = b_n \to 0$, we use $\mathcal{O}(b^{1/\gamma_1})$ observations with $x < 0$ and $\mathcal{O}(b^{1/\gamma_2}) \gg \mathcal{O}(b^{1/\gamma_1})$ observations with $x > 0$ (assuming x_i's behave as if sampled from positive, bounded density).

- In particular, if all x_i are positive, then $\hat{L}_{\omega^{-1}(b),\mathcal{F},\mathcal{G}} = L_{\omega^{-1}(2b),\mathcal{F},\mathcal{F}}$. If all x_i are negative, then $\hat{L}_{\omega^{-1}(b),\mathcal{F},\mathcal{G}} = L_{\omega^{-1}(2b),\mathcal{G},\mathcal{G}}$.

- Thus, there are potential gains from adaptation when 0 is a left boundary of supp(x), but not when it is a right boundary. There is an intuitive reason for this: if we know that f is nonincreasing, then we can get a lower-biased estimate of $f(0)$ using observations with $x_i > 0$, which gives a lower CI. However, we cannot do this with observations where $x_i < 0$.

3.4.2 Comparison with other "nontestability" results

- Consider $\{X_i\}_{i=1}^n$ iid with $X_i \sim F$ where F is known to be in the parameter space

  $$\mathcal{F}(C) = \{\text{distributions on } \mathbb{R} \text{ with absolute 3rd moment bounded by } C\},$$

  and let the parameter of interest be the mean

  $$LF = E_F X = \int x \, dF(x).$$

- It follows from Bahadur and Savage (1956) that any $1-\alpha$ CI $[\hat{c}, \infty)$ for LF that is valid over $\mathcal{F} = \cup_{C=0}^{\infty} \mathcal{F}(C)$ (i.e. the set of distributions for which the 3rd moment exists) must be trivial in the sense that it satisfies $q_{\beta,F}(LF - \hat{c}) = \infty$ for all $F \in \mathcal{F}$ and $\beta > \alpha$.

- However, if we pick $\overline{C}$ and only require coverage over $\mathcal{F}(\overline{C})$, we can get a CI (using, e.g. Berry-Esseen) that is "adaptive" to the variance of F in the sense that $\sqrt{n} q_{\beta,F}(LF - \hat{c}) \to (z_\beta + z_{1-\alpha})\sigma(F)$ for all F with $\sigma^2(F) > \underline{\sigma}^2 > 0$ (which achieves the semiparametric efficiency bound).

- The CI takes the form $\bar{X} - z_{1-\alpha}(\hat{\sigma}/\sqrt{n}) \cdot (1 + \eta_n(\overline{C}))$ where $\eta_n(\overline{C}) \to 0$. Thus, while the CI does need to depend explicitly on $\overline{C}$ for finite sample validity, the dependence is second-order.

- In contrast, the lack-of-adaptation results for adapting to C or γ in, say, a Hölder class, show that the optimal $q_{\beta,f}(Lf - \hat{c})$ depends explicitly on $\overline{C}$ even asymptotically.

References

Armstrong, T. B. (2014): "Weighted KS statistics for inference on conditional moment inequalities," *Journal of Econometrics*, 181(2), 92–116.

Armstrong, T. B., and M. Kolesár (2015): "Optimal inference in a class of regression models," *arXiv:1511.06028 [math, stat]*.

Bahadur, R. R., and L. J. Savage (1956): “The Nonexistence of Certain Statistical Procedures in Nonparametric Problems,” *The Annals of Mathematical Statistics*, 27(4), 1115–1122.

Beliakov, G. (2006): “Interpolation of Lipschitz functions,” *Journal of Computational and Applied Mathematics*, 196(1), 20–44.

Brown, L. D., and M. G. Low (1996): “Asymptotic equivalence of nonparametric regression and white noise,” *The Annals of Statistics*, 24(6), 2384–2398.

Cai, T. T., and M. G. Low (2004): “An Adaptation Theory for Nonparametric Confidence Intervals,” *The Annals of Statistics*, 32(5), 1805–1840.

Cheng, M.-Y., J. Fan, and J. S. Marron (1997): “On automatic boundary corrections,” *The Annals of Statistics*, 25(4), 1691–1708.

Cook, T. D. (2008): ““Waiting for Life to Arrive”: A history of the regression-discontinuity design in Psychology, Statistics and Economics,” *Journal of Econometrics*, 142(2), 636–654.

Donoho, D., and J. Jin (2004): “Higher criticism for detecting sparse heterogeneous mixtures,” *The Annals of Statistics*, 32(3), 962–994.

Donoho, D. L. (1994): “Statistical Estimation and Optimal Recovery,” *The Annals of Statistics*, 22(1), 238–270.

Donoho, D. L., and M. G. Low (1992): “Renormalization Exponents and Optimal Pointwise Rates of Convergence,” *The Annals of Statistics*, 20(2), 944–970.

Elliott, G., U. K. Müller, and M. W. Watson (2015): “Nearly Optimal Tests When a Nuisance Parameter Is Present Under the Null Hypothesis,” *Econometrica*, 83(2), 771–811.

Ingster, Y., and I. A. Suslina (2003): *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Springer.

Johnstone, I. M. (2015): *Gaussian estimation: Sequence and wavelet models*. Online manuscript available at http://statweb.stanford.edu/~imj/.

Lehmann, E. L. (1952): "Testing Multiparameter Hypotheses," *The Annals of Mathematical Statistics*, 23(4), 541–552.

Lehmann, E. L., and J. P. Romano (2005): *Testing statistical hypotheses*. Springer.

Lepski, O., and A. Tsybakov (2000): "Asymptotically exact nonparametric hypothesis testing in sup-norm and at a fixed point," *Probability Theory and Related Fields*, 117(1), 17–48.

Low, M. G. (1995): "Bias-Variance Tradeoffs in Functional Estimation Problems," *The Annals of Statistics*, 23(3), 824–835.

Nussbaum, M. (1996): "Asymptotic equivalence of density estimation and Gaussian white noise," *The Annals of Statistics*, 24(6), 2399–2430.

Romano, J. P., A. M. Shaikh, and M. Wolf (2013): "A Practical Two-Step Method for Testing Moment Inequalities," *Unpublished Manuscript*.

Sacks, J., and D. Ylvisaker (1978): "Linear Estimation for Approximately Linear Models," *The Annals of Statistics*, 6(5), 1122–1137.

——— (2012): "After 50+ Years in Statistics, An Exchange," *Statistical Science*, 27(2), 308–318.

Thistlethwaite, D. L., and D. T. Campbell (1960): "Regression-discontinuity analysis: An alternative to the ex post facto experiment," *Journal of Educational Psychology*, 51(6), 309–317.