STATS 300C Notes

Kenneth Tay

Contents

1 Global Testing (Lec 1)

(Lec 1) We define the **global null** $H_0 = \bigcap_{i=1}^{n} H_{0,i}$.

1.1 Bonferroni's method (Lec 1-2)

Let p_i be the p-value for testing $H_{0,i}$. For a level α test: reject whenever $\min_i p_i \leq \alpha/n$.

- (Lec 1) Size of test (i.e. probability of type 1 error under null) is $\leq \alpha$. If the hypotheses are independent, then size $\mathbb{P}_{H_0}(\text{Type I error}) \xrightarrow{n\to\infty} 1 - e^{-\alpha}$.
- Most suited for cases where we expect at least one of the p-values to be very significant.
- (Lec 2) When we are looking at $y_i \overset{ind}{\sim} \mathcal{N}(\mu_i, 1)$ and testing $H_{0,i} : \mu_i = 0$, Bonferroni rejects if $\max y_i \geq |z(\alpha/n)|$ in the one-sided case, and if $\max y_i \geq |z(\alpha/2n)|$ in the two-sided case.
- (Lec 2) Holding α fixed, for large n,

$$|z(\alpha/n)| \approx \sqrt{2\log n}\left[1 - \frac{\log\log n}{4\log n}\right] \approx \sqrt{2\log n}, \qquad \frac{\max|y_i|}{\sqrt{2\log n}} \xrightarrow{P} 1.$$

  No dependence on α!

- (Lec 2) More accurate approximation: If $B = 2\log(n/\alpha) - \log(2\pi)$, then $|z(\alpha/2n)| \approx \sqrt{B\left(1 - \frac{\log B}{B}\right)}$.

- (Lec 2) **Needle in a haystack problem:** Suppose the alternative is such that exactly one $\mu_i = \mu^{(n)} > 0$ (we don't know which one).

  1. If $\mu^{(n)} = (1+\varepsilon)\sqrt{2\log n}$, then Bonferroni has asymptotic full power.
  2. If $\mu^{(n)} = (1-\varepsilon)\sqrt{2\log n}$, then Bonferroni has asymptotic powerlessness, i.e. $\mathbb{P}_{H_1}(\max y_i > |z(\alpha/n)|) \to \alpha$.
  3. If we use $\sqrt{2\log n}$ instead of $|z(\alpha/n)|$ as the threshold for the test, then we can achieve $\mathbb{P}_{H_0}(\text{Type I Error}) \to 0$ and $\mathbb{P}_{H_1}(\text{Type II Error}) \to 0$.

- (Lec 2) **Optimality of Bonferroni:** When $\mu^{(n)} = (1-\varepsilon)\sqrt{2\log n}$, no test can do better than Bonferroni. (Proof of optimality sets up an easier "Bayesian" decision problem and shows the optimality for that set-up.)

1.2 Fisher's combination test (Lec 1)

Reject for large values of $T = -\sum_{i=1}^{n} 2\log p_i$.

- If $p_i \overset{iid}{\sim} \text{Unif}(0,1)$ under the null, then under the null, $T \sim \chi^2_{2n}$. Thus, Fisher's test rejects when $T > \chi^2_{2n}(1-\alpha)$.

- Most suited where we expect many small effects.

1.3 χ^2 test (Lec 3)

Model $Y \sim \mathcal{N}(\mu, I)$. Testing $\mu = 0$ vs. $\mu \neq 0$. Test statistic is $T = \|y\|^2 = \sum_{i=1}^{n} y_i^2$.

- Under the null, $T \sim \chi^2_n$, so we reject when $T > \chi^2_n(1-\alpha)$.

- Let $Z = \dfrac{T-n}{\sqrt{2n}}$ be the normalized version of the statistic, and let $\theta = \dfrac{\|\mu\|^2}{\sqrt{2n}}$. Then under H_0, $Z \sim \mathcal{N}(0,1)$ and under H_1, $Z \sim \mathcal{N}\left(\theta, 1 + \dfrac{\theta}{\sqrt{n/8}}\right)$.

  Roughly speaking, the test is easy when θ is large and difficult when θ is small. Thus, the power of the χ^2 test is determined by the relative size of $\|\mu\|^2$ compared to $\sqrt{n}$.

- θ can be thought of as the **signal-to-noise ratio**.

- When $\theta \ll 1$ ($\theta \to 0$), the χ^2 test is asymptotically powerless, but so are all other tests. (Proof uses a simpler Bayesian decision problem and shows the optimal test there is powerless.)

- As with the Fisher combination test, the χ^2 test is powerful when there are many small, distributed effects but weak when there are few strong effects.

1.4 Simes test (Lec 4)

Assume that we have p-values $p_i \sim \text{Unif}(0,1)$. Order them: $p_{(1)} \le \ldots \le p_{(n)}$. The **Simes statistic** is $T_n = \min_i \left\{ p_{(i)} \frac{n}{i} \right\}$ (i.e. smaller ones get inflated more).

- Under the global null and independence of the p_i, $T_n \sim \text{Unif}(0,1)$. (Proof by induction.) Hence, the test rejects if $T_n \le \alpha$. (Equivalently, the test rejects if there is an i such that $p_{(i)} \le \dfrac{\alpha i}{n}$.
- Simes test still has level α under some sort of positive dependence (PRDS).
- Simes test is strictly less conservative than Bonferroni, which rejects for $p_{(1)} \le \alpha/n$.
- Simes test is powerful for a single strong effect, but has moderate power for many mild effects.

1.5 Tests based on empirical CDFs (Lec 4)

The **empirical CDF** of $p_1, \ldots, p_n$ is $\widehat{F}_n(t) = \dfrac{\#\{i : p_i \le t\}}{n}$.

Under the global null H_0, we have $\mathbb{E}[\widehat{F}_n(t)] = t$. If we further assume that the p_i's are independent, we have that $n\widehat{F}_n(t) \sim \text{Binom}(n,t)$.

- **Kolmogorov-Smirnov test:** $KS = \sup_t |\widehat{F}_n(t) - t|$. Reject if KS rejects a certain threshold (which can be computed through simulation or asymptotic calculation).
- **Anderson-Darling test:** Let $w(t)$ be a non-negative weight function. Define $A = n \displaystyle\int_0^1 \left(\widehat{F}_n(t) - t\right)^2 w(t)dt$.
  - When $w(t) = 1$, A^2 is the **Cramer-von Mises statistic**.
  - When $w(t) = \dfrac{1}{t(1-t)}$, A^2 is the **Anderson-Darling statistic**. It puts more weight on small/large p-values than the Cramer-von Mises statistic.
  - We can write the Anderson-Darling statistic as

    $$A^2 = -n - \sum_{i=1}^n \frac{2i-1}{n}[\log(p_{(i)}) + \log(1 - p_{(n+1-i)})].$$

    It gives more weight to p-values in the bulk than Fisher's combination statistic.
- **Tukey's second-Level significance testing: Higher criticism statistic** $HC_n^* = \max\limits_{0 \le t \le \alpha_0} \dfrac{\widehat{F}_n(t) - t}{\sqrt{t(1-t)/n}}$.

1.6 Sparse mixtures (Lec 4)

$H_0 : X_i \overset{iid}{\sim} \mathcal{N}(0,1)$, $H_1 : X_i \overset{iid}{\sim} (1-\varepsilon)\mathcal{N}(0,1) + \varepsilon\mathcal{N}(\mu,1)$. If we set fraction of non-nulls $\varepsilon_n = n^{-\beta}$ for some $\beta \in (1/2, 1)$ and $\mu_n = \sqrt{2r \log n}$ for some $r \in (0,1)$, then there is a threshold curve

$$\rho^*(\beta) = \begin{cases} \beta - 1/2 & \text{if } \frac{1}{2} \le \beta \le \frac{3}{4}, \\ (1 - \sqrt{1-\beta})^2 & \text{if } \frac{3}{4} \le \beta \le 1, \end{cases}$$

such that

1. If $r > \rho^*(\beta)$, we can adjust the NP test to achieve $P_0(\text{Type I error}) + P_1(\text{Type II error}) \to 0$.
2. If $r < \rho^*(\beta)$, then for **any** test, $\liminf_n P_0(\text{Type I error}) + P_1(\text{Type II error}) \geq 1$.

For $r > \rho^*(\beta)$, the higher criticism statistic (with an appropriate threshold) has full power asymptotically.

2 Multiple Testing: Family-wise error rate (FWER) (Lec 5)

We have 4 types of outcomes in multiple testing:

	not rejected	rejected	total
true nulls	U	V	n_0
true non-nulls	T	S	$n - n_0$
total	$n - R$	R	n

Only n and R are observed random variables; U, V, S, T, n_0 are all unobserved.

Family-wise error rate is defined as the probability of at least 1 false rejection, i.e. $FWER = \mathbb{P}(V \geq 1)$.

A procedure controls FWER **strongly** if FWER is controlled under all configurations of true and false hypotheses. It controls FWER **weakly** if FWER is controlled under the global null.

2.1 Bonferroni's method (Lec 5)

Reject all $H_{0,i}$ for which $p_i \leq \alpha/n$.

- Bonferroni's method controls FWER strongly at level α (even when hypotheses are dependent).
- **Sidak's procedure:** Under independence, we can reject all $H_{0,i}$ for which $p_i \leq \alpha_n$, where α_n is slightly bigger than α/n.

2.2 Fisher's two-step procedure (Lec 5)

First, do a test for the global null. If not rejected, stop. If rejected, then test each hypothesis at level α.

This procedure only controls FWER weakly.

2.3 Holm's procedure (Lec 5)

Holm's procedure is a step-down procedure (from most significant p-value to least significant p-value). Order the p-values $p_{(1)} \leq \ldots \leq p_{(n)}$, and let $H_{(i)}$ be the hypothesis corresponding to $p_{(i)}$.

1. Step 1: If $p_{(1)} \leq \alpha/n$, reject $H_{(1)}$ and go to step 2. Otherwise, stop and "accept" $H_{(1)}, \ldots, H_{(n)}$.

2. Step i: If $p_{(i)} \leq \alpha/(n-i+1)$, reject $H_{(i)}$ and go to step $i+1$. Otherwise, stop and "accept" $H_{(i)}, \ldots, H_{(n)}$.

Basically, stop the first time $p_{(i)}$ exceeds $\alpha_i = \alpha/(n-i+1)$ (reject everything less than i, "accept" everything $\geq i$).

Holm's procedure controls FWER strongly.

2.4 Closure principle (Lec 6)

- For $\{H_i\}_{i=1}^n$, we can define its **closure** to be $H_I = \bigcap_{i \in I} H_i$ for $I \subseteq \{1, \ldots, n\}$.
- **Closure procedure:** Reject H_I iff for all $J \supseteq I$, H_J is rejected at level α.
- **Theorem:** Closing a global test gives a procedure which controls FWER strongly.
- **Closing Bonferroni gives Holm's procedure.**

2.5 Hochberg's procedure (Lec 6)

Hochberg's procedure is a step-up procedure (from least significant p-value to most significant p-value). Order the p-values $p_{(1)} \leq \ldots \leq p_{(n)}$, and let $H_{(i)}$ be the hypothesis corresponding to $p_{(i)}$.

Hochberg procedure: Reject $H_{(j)}$ if there is an index $j' \geq j$ such that $p_{(j')} \leq \dfrac{\alpha}{n-j'+1}$.

- Hochberg scans backwards, and stops as soon as a p-value succeeds in passing its threshold.
- **Hochberg's procedure is more conservative than the closure of Simes.**
- Hochberg's procedure requires independence of null p-values.

3 Multiple Testing: False discovery rate (FDR) (Lec 7)

We have 4 types of outcomes in multiple testing:

	not rejected	rejected	total
true nulls	U	V	n_0
true non-nulls	T	S	$n - n_0$
total	$n-R$	R	n

Only n and R are observed random variables; U, V, S, T, n_0 are all unobserved.

False discovery proportion (FDP) is defined to be

$$FDP = \frac{V}{\max(R,1)} = \begin{cases} V/R & \text{if } R \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

FDP is an unobserved variable. Instead, we control its expectation, the **false discovery rate:** $FDR = \mathbb{E}[FDP]$.

- Under the global null, FDR is equivalent to FWER. Thus, FDR control implies weak FWER control.
- $FWER \geq FDR$. Thus, controlling FWER (strongly) implies FDR control.
- Alternative to FDR: **false exceedence rate:** $\mathbb{P}(FDP \geq q)$.
- $1\{V \geq 1\} \geq FDP$.

3.1 Benjamini-Hochberg (BHq) procedure (Lec 7)

Fix some level $q \in [0,1]$. Order the p-values: $p_{(1)} \leq \ldots \leq p_{(n)}$. Let i_0 be the largest i for which $p_{(i)} \leq \frac{iq}{n}$. Reject all $H_{(i)}$ with $i \leq i_0$.

- **Thm:** For independent test statistics (p-values), the BHq procedure controls the FDR at level q. In fact, $FDR = \frac{n_0 q}{n} \leq q$, where n_0 is the number of true nulls.
  - Lec 7 has a proof where we write $FDP = \sum_{i \in H_0} \frac{V_i}{1 \vee R}$, where H_0 is the set of true nulls and $V_i = 1\{H_i \text{ rejected}\}$. To make the denominator tractable, rewrite as $\frac{V_i}{1 \vee R} = \sum_{k=1}^{n} \frac{V_i 1\{R = k\}}{k}$.
  - From the proof in Lec 7, we see that we only need independence of the true null p-values among themselves and from the non-nulls. We do not require independence between the non-null p-values.
  - Lec 8 has a martingale proof of the FDR control of BHq, where the Optional Stopping Theorem is used for $V(t)/t$, the martingale running backward in time.
- Under the global null, BHq is Simes.
- BHq, like Hochberg's procedure, is a step-up procedure. BHq is approximately i times more liberal than Hochberg's procedure (for small values of i). This is seen by comparing the ratio of the thresholds.
- (HW2) BHq at level q does NOT control FWER at level q. BHq at level q does control FWER at level $nq \wedge 1$.

3.1.1 Empirical process viewpoint of BHq (Lec 8)

Take a fixed t, and consider the rule that rejects hypothesis H_i iff $p_i \leq t$. Then we have

	not rejected	rejected	total
true nulls	$U(t)$	$V(t)$	n_0
true non-nulls	$T(t)$	$S(t)$	$n - n_0$
total	$n - R(t)$	$R(t)$	n

We have $FDP(t) = \frac{V(t)}{1 \vee R(t)}$ and $FDR(t) = \mathbb{E}\left[\frac{V(t)}{1 \vee R(t)}\right]$.

- If we have an estimate $\widehat{FDR}(t)$ for $FDR(t)$, we can take the threshold $\tau = \sup\{t \leq 1 : \widehat{FDR}(t) \leq q\}$. This defines the most liberal thresholding cut-off.

- A conservative estimate takes $\mathbb{E}[V(t)] = n_0 t \leq nt$. This gives $\widehat{FDR}(t) = \dfrac{nt}{1 \vee R(t)} = \dfrac{t}{\hat{F}_n(t) \vee 1/n}$.
- **The above is exactly what BHq is doing.** We stop at $p^* = \max\left\{t \in \{p_1, \ldots, p_n\} : \dfrac{t}{q} \leq \hat{F}_n(t)\right\}$. Thus, if we write

$$\tau_{BH} = \max\left\{t : \frac{t}{\hat{F}_n(t) \vee 1/n} \leq q\right\},$$

then the BH procedure rejects all hypotheses with $p_i \leq \tau_{BH}$. We also have $\tau_{BH} \geq q/n$.

- To improve on BHq, we can try a less conservative estimate of $\widehat{FDR}(t)$, which amounts to a less conservative estimate of $\pi_0 = \dfrac{n_0}{n}$ (fraction of true nulls). For example, we could try $\pi_0^\lambda = \dfrac{n - R(\lambda)}{(1-\lambda)n}$.

3.1.2 BHq under dependence (Lec 9)

- There are joint distributions of p-values for which the FDR of the BHq procedure is at least $q \cdot S(n) \wedge 1$, where $S(n) = 1 + \dfrac{1}{2} + \dfrac{1}{3} + \cdots + \dfrac{1}{n} \approx \log n + 0.577$.
- More generally, if we have a BH(α) procedure where the critical values are $0 \leq \alpha_1 \leq \ldots \leq \alpha_n \leq 1$, there is a joint distribution of p-values for which the FDR of the BHq procedure is at least $\left(\sum_{k=1}^{n} \dfrac{n(\alpha_k - \alpha_{k-1})}{k}\right) \wedge 1$.
- On the flip-side, we can show that under dependence, the BHq procedure controls FDR at level $q \cdot S(n)$. In fact, $FDR \leq q \cdot S(n) \cdot \dfrac{n_0}{n}$. (Proof uses the trick of decomposing FDP into a sum of $\dfrac{V_i}{1 \vee R}$.)
- A set $D \in \mathbb{R}^n$ is **increasing** if $x \in D$ and $y \geq x$ (component-wise) implies $y \in D$. (These sets have no boundaries in the northeast directions.)
- A family of random variables $(X_1, \ldots, X_n)$ is **PRDS (positive regression dependence on subset)** on I_0 if for any increasing set $D \in \mathbb{R}^n$ and each $i \in I_0$, $\mathbb{P}((X_1, \ldots, X_n) \in D \mid X_i = x)$ is an increasing function of x.
  - The PRDS property is invariant by co-monotone transformations: If $Y_i = f_i(X_i)$, where all the f_i's are either increasing or decreasing, then X is PRDS implies that Y is PRDS.
  - D is increasing iff D^c is decreasing. Thus, X is PRDS iff for any decreasing C, $\mathbb{P}(X \in C \mid X_i = x)$ is decreasing in x.
  - Because CDFs are monotone, if test statistics $\{X_i\}$ are PRDS on I_0 (set of true nulls), then the one-sided p-values are both PRDS. However, the two-sided p-values may not be PRDS.
  - **Multivariate normal:** Let $X = (X_1, \ldots, X_n) \sim \mathcal{N}(\mu, \Sigma)$. If $\Sigma_{ij} \geq 0$ for all $i \in I_0$ and all j, then X is PRDS on I_0. (The converse also holds.)
- If the joint distribution of the p-values/statistics is PRDS on the set of true nulls, then BHq controls the FDR at level $\dfrac{qn_0}{n}$. (Note: BHq may become conservative under positive dependence.) (Proof uses the usual decomposition of FDP and integration by parts.)
  - In this setting, if i is a null and D is an increasing set, then for $t \leq t'$, we have $\mathbb{P}(D \mid p_i \leq t) \leq \mathbb{P}(D \mid p_i \leq t')$.

3.2 Bayesian FDR (Lec 11)

- We have n hypotheses, which are null ($H = 0$) with probability π_0, and non-null ($H = 1$) with probability $\pi_1 = 1 - \pi_0$.
- Let our test statistics be $z \sim f_0$ with CDF F_0 if $H = 0$, and $z \sim f_1$ with CDF F_1 if $H = 1$. (Marginally, $f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$.)
- For a set A, let $F_0(A) = \displaystyle\int_A f_0(z)dz$, $F(A) = \displaystyle\int_A f(z)dz$.
- Think of A as a rejection region. If I observe $z \in A$, I will reject the corresponding hypothesis. In practice, A is of one of the following forms: $[z_c, \infty)$, $(-\infty, -z_c]$ or $(-\infty, -z_c] \cup [z_c, \infty)$.
- **Bayes false discovery rate (BFDR)** is defined to be $\varphi(A) = \mathbb{P}(H = 0 \mid z \in A) = \dfrac{\pi_0 F_0(A)}{F(A)}$. (If we report $z \in A$ as a non-null, $\varphi(A)$ is the probability that we've made a false discovery.)
- We can distinguish between global BFDR ($\varphi((-\infty, z_c])$) and local BFDR $\varphi(\{z_c\})$. See Lec 11 for details.

3.2.1 Empirical Bayes estimation of BFDR

- To compute BFDR, we need to know π_0, F_0 and F. In the following, we assume that f_0 is known (and assumed to be $\mathcal{N}(0, 1)$), $\pi_0 \approx 1$ (true for most applications), and f_1 is unknown.
- Estimate $\widehat{F}(A) = \dfrac{\#\{z_i \in A\}}{n}$, $\widehat{BFDR} = \dfrac{\pi_0 F_0(A)}{\widehat{F}(A)}$.
- Let $N_0(A) = \#\{i : H_i \text{ is true null and } z_i \in A\}$, $N_+(A) = \#\{i : z_i \in A\}$, $e_0(A) = \mathbb{E}N_0(A)$, $e_+(A) = \mathbb{E}N_+(A)$. In this notation,

$$BFDR = \frac{e_0(A)}{e_+(A)}, \qquad \widehat{BFDR} = \frac{e_0(A)}{N_+(A)}.$$

- **Lemma:** If we let $\gamma(A)$ denote the squared coefficient of variation of $N_+(A)$, i.e. $\gamma(A) = \dfrac{\text{Var } N_+(A)}{[e_+(A)]^2}$, then $\dfrac{\widehat{BFDR}}{BFDR}$ has mean approximately $1 + \gamma$ and variance γ.
- $\widehat{BFDR}$ is a reasonably accurate estimator if e_+ is large.
- The empirical Bayes formulation of BHq is to reject H_i for all $i \leq i_0$, where i_0 is the largest index such that $\widehat{BFDR}((-\infty, z_{(i_0)}]) \leq q$. Assuming independence of statistics, the FDR is at most q.

4 Knockoffs (Lec 12)

4.1 Regression setting (Lec 12)

We have a linear model $Y = X\beta + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, I)$. We wish to control the FDR when testing $H_j : \beta_j = 0$, $j = 1, \ldots, p$.

Suppose we have the full lasso path, i.e. for each λ we compute $\hat{\beta}(\lambda) = \text{argmin}_{b \in \mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \lambda\|b\|_1$. We look at when each β_j enters the lasso path: $Z_j = \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$. Our idea is to reject all H_j for which $Z_j \geq T$. We need some principled way to determine the threshold T.

- We create **knockoffs** with 3 requirements, the first two being the most important (See Lec 12 for construction):
  1. For all j, k, $\widetilde{X}_j' \widetilde{X}_k = X_j' X_k$.
  2. For all $j \neq k$, $\widetilde{X}_j' X_k = X_j' X_k$.
  3. For all j, $\widetilde{X}_j' X_j$ is as small as possible.
- The knockoffs are not unique.
- **Pairwise exchangeability:** For any subset S of nulls, $\begin{bmatrix} X & \widetilde{X} \end{bmatrix}_{swap(S)}' y \overset{d}{=} \begin{bmatrix} X & \widetilde{X} \end{bmatrix}' y$, where $\begin{bmatrix} X & \widetilde{X} \end{bmatrix}_{swap(S)}$ means that columns X_j and $\widetilde{X}_j$ have been swapped for every $j \in S$.
- Compute the lasso estimates for the regression $y = X\beta + \widetilde{X}\widetilde{\beta} + \varepsilon$. Look at where the original and knockoff variables first enter the lasso path, i.e. $Z_j = \sup\{\lambda : \beta_j(\lambda) \neq 0\}$, $\widetilde{Z}_j = \sup\{\lambda : \widetilde{\beta}_j(\lambda) \neq 0\}$.
- Consider the test statistic (for hypothesis H_j)

  $$W_j = \max(Z_j, \widetilde{Z}_j) \cdot \text{sign}(Z_j - \widetilde{Z}_j).$$

  If W_j is large and positive, it provides good evidence that X_j is non-null.
- **Theorem 1:** Given a target FDR q, reject all Z_j which are greater than or equal to

  $$T = \min\left\{t : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\}.$$

  This gives $\mathbb{E}\left[\frac{V}{R + q^{-1}} \leq q\right]$. (Note: The fraction in the stopping time is the estimate $\widehat{FDP}(t)$.)
- **Theorem 2:** Given a target FDR q, reject all Z_j which are greater than or equal to

  $$T = \min\left\{t : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\}.$$

  We get exact FDR control: $\mathbb{E}\left[\frac{V}{R \vee 1} \leq q\right]$.
- See HW4 Qn1 for the general technique of proof.

4.2 Model-free knockoffs (Lec 13)

Model assumptions:

- We have n samples $(X^{(i)}, Y^{(i)})$ i.i.d. sampled from some joint distribution F_{XY}.
- The distribution F_X of X is known.

- The conditional distribution $F_{Y|X}$ of Y given X is completely unknown.

We construct knockoff variables $X = (\widetilde{X}_1, \ldots, \widetilde{X}_p)$ such that for any subset $\mathcal{T} \subseteq \{1, 2, \ldots, p\}$, we have $(X, \widetilde{X})_{swap(\mathcal{T})} \overset{d}{=} (X, \widetilde{X})$. (That is, swapping any subset of variables with their knockoffs does not change the joint distribution.) We also require that the knockoffs be constructed without any knowledge of Y.

- **Multivariate normal:** If $X \sim \mathcal{N}(\mu, \Sigma)$, we can simply sample $\widetilde{X}$ so that the joint distribution of X and $\widetilde{X}$ is
$$\begin{pmatrix} X \\ \widetilde{X} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} \Sigma & \Sigma - \operatorname{diag}(s) \\ \Sigma - \operatorname{diag}(s) & \Sigma \end{pmatrix} \right),$$
where s is chosen so that the covariance matrix is positive semidefinite.
- (Lec 14) **General construction (Sequential Conditional Independent Pairs (SCIP):** Let $j = 1$. While $j \leq p$, sample $\widetilde{X}_j$ from the law of $X_j \mid X_{-j}, \widetilde{X}_{1:j-1}$, and increment j. (In general, not a practical algorithm.)
- See Lec 14 for construction of knockoffs for Markov chains and hidden Markov models.

After constructing the knockoffs, we run a procedure on the original with the knockoff variables serving as controls (no assumption on what the procedure is). From this procedure, we get Z_i for each variable which signals the importance of variable i in the model. We also construct these important statistics $\widetilde{Z}_i$ for the knockoff features.

- If we write $(Z_1, \ldots, Z_p, \widetilde{Z}_1, \ldots, \widetilde{Z}_p) = z((X, \widetilde{X}), Y)$, then we require that swapping the original variables with their knockoffs simply swaps the test statistics, i.e. $(Z, \widetilde{Z})_{swap(\mathcal{T})} = z((X, \widetilde{X})_{swap(\mathcal{T})}, Y)$ for any subset $\mathcal{T}$.
- **Theorem:** For any subset $T \subseteq \mathcal{H}_0$ of nulls, we always have $(Z, \widetilde{Z})_{swap(T)} \overset{d}{=} (Z, \widetilde{Z})$.

Next, we combine Z_j and $\widetilde{Z}_j$ into a single test statistic W_j, i.e. $W_j = w_j(Z_j, \widetilde{Z}_j)$. We require that w_j is **anti-symmetric**.

- We have an estimate for FDP: $\widehat{FDP}(t) = \dfrac{\#\{W_j \leq -t\}}{\#\{W_j \geq t\} \vee 1}$.
- **Theorem:** Set $N^{\pm}(t) = \#\{j : |W_j| \geq t \text{ and } \operatorname{sign}(W_j) = \pm\}$, $T_{0/1} = \min \left\{ t : \widehat{FDP}(t) = \dfrac{0/1 + N^-(t)}{1 \vee N^+(t)} \leq q \right\}$.
Select variables $\hat{\mathcal{S}} = \{W_j > T\}$.
With $T = T_0$, we have $\mathbb{E}\left[\dfrac{V}{R + q^{-1}} \leq q \right]$. With $T = T_1$, we have exact FDR control: $\mathbb{E}\left[\dfrac{V}{R \vee 1} \right] \leq q$.

5 Selective Inference (Lec 15)

Say we have n parameters $\theta_1, \ldots, \theta_n$ with corresponding statistics $T_1, \ldots, T_n$. Assume that we have α-level confidence intervals $CI_i(\alpha)$ for each θ_i.

- **Marginal coverage:** $\mathbb{P}(\theta_i \in CI_i(\alpha)) \geq 1 - \alpha$.

- **Simultaneous coverage:** $\mathbb{P}((\theta_1, \ldots, \theta_n) \in CI(\alpha)) \geq 1-\alpha$. (This can be achieved by doing Bonferroni correction on Wald intervals.)

Say we have selected a subset $\mathcal{S}$ of parameters. Then marginal confidence intervals (e.g. Wald intervals) will not have the desired coverage. **Conditional coverage** is $\mathbb{P}_\theta(\theta_i \in CI_i(\alpha) \mid i \in \mathcal{S}) \geq 1-\alpha$. (In general cannot be achieved.)

5.1 False coverage rate (Lec 15)

- Define **false coverage rate** to be $FCR = \mathbb{E}\left[\frac{V_{CI}}{R_{CI} \vee 1}\right]$, where R_{CI} is the number of selected parameters and V_{CI} is the number of constructed intervals not covering the parameter (out of the selected ones).
  - Without selection, the marginal CIs control FCR.
  - Bonferroni CIs do control FCR (in the same way Bonferroni's procedure controls FDR).
- Consider the following procedure:
  1. Apply some subsection rule $\mathcal{S}(T)$, where T are the statistics $T_1, \ldots, T_n$ for parameters $\theta_1, \ldots, \theta_n$.
  2. For each $i \in \mathcal{S}$, let $R_{min}(T^{(i)}) = \min_t \left\{|\mathcal{S}(T^{(i)}, t)| \colon i \in \mathcal{S}(T^{(i)}, t)\right\}$, where $T^{(i)} = T \setminus \{T_i\}$. That is, the size of the smallest possible set which can be selected such that (i) it still selects i, and (ii) statistics $T^{(i)}$ are fixed as before.
  3. For each $i \in \mathcal{S}$, the FCR-adjusted confidence interval is $CI_i\left(\frac{R_{min}(T^{(i)})\alpha}{n}\right)$.

  If the T_i's are independent, then for any selection procedure, the FCR of the adjusted CI's obey $FCR \leq \alpha$.

5.2 Post selection inference (POSI) and selective inference for LASSO

See Lec 16 for details.

5.3 Selective hypothesis testing

See Lec 17 for details.

6 Estimation of Multivariate Normal Mean (Lec 18)

Consider the problem of estimating μ in the model $X \sim \mathcal{N}(\mu, \sigma^2 I)$ under loss function $\ell(\mu, \hat{\mu}) = \|\mu - \hat{\mu}\|^2$ (i.e. squared error loss).

- The most natural estimator, also the MLE, is X itself. This estimator has constant risk $R(\hat{\mu}_{MLE}, \mu) = p\sigma^2$.
- For $p = 1, 2$, the MLE is admissible. However, it is inadmissible for $p \geq 3$!

- **James-Stein estimator:** $\hat{\mu}_{JS} = \left(1 - \frac{(p-2)\sigma^2}{\|X\|^2}\right) X$. It is non-linear, biased, and shrinks the MLE towards 0. It dominates the MLE everywhere in terms of MSE.
  - $\hat{\mu}_{JS}$ is itself inadmissible: $\hat{\mu}_{JS}^{+} = \left(1 - \frac{(p-2)\sigma^2}{\|X\|^2}\right)_{+} X$ dominates it.
- **Stein's unbiased risk estimate (SURE):** Suppose $X \sim \mathcal{N}(\mu, \sigma^2 I)$, and that estimator $\hat{\mu} = X + g(X)$, where g is almost-differentiable, and $\mathbb{E}\left[\sum_{i=1}^{p} |\partial_i g_i(X)|\right] < \infty$. (Almost-differentiable means there exist h_i such that $g_i(x+z) - g_i(x) = \int_0^1 \langle h_i(x+tz), z\rangle dt$. We usually write $h_i = \nabla g_i$.) Then

$$\mathbb{E}\|\mu - \hat{\mu}\|^2 = p\sigma^2 + \mathbb{E}\left[\|g(X)\|^2 + 2\sigma^2 \sum_i \partial_i g_i(X)\right],$$
$$SURE(\hat{\mu}) = p\sigma^2 + \|g(X)\|^2 + 2\sigma^2 \text{div } g(X),$$

  i.e. we have an expression which is unbiased for the estimator's risk.

6.1 Empirical Bayes interpretation (Lec 19)

Consider the Bayes model $\mu_i \overset{iid}{\sim} \mathcal{N}(0, \tau^2)$, $X \mid \mu \sim \mathcal{N}(\mu, \sigma^2 I)$.

- Posterior distribution is $\Lambda(\mu \mid X) \sim \mathcal{N}\left(\frac{\tau^2}{\tau^2 + \sigma^2} X, \frac{\tau^2 \sigma^2}{\tau^2 + \sigma^2} I\right)$, so Bayes estimate would be $\hat{\mu}_B = \left(1 - \frac{\sigma^2}{\sigma^2 + \tau^2}\right) X$.
- Bayes risk can be computed to be $\mathbb{E}\|\hat{\mu}_B - \mu^2\| = R_{MLE} \frac{\tau^2}{\tau^2 + \sigma^2}$. (See Lec 19 for details.)
- In practice, we don't know τ. We can estimate τ from the fact that $\|X\|^2 \sim (\tau^2 + \sigma^2)\chi_p^2$, so an unbiased estimate for $\frac{\sigma^2}{\sigma^2 + \tau^2}$ is $\frac{(p-2)\sigma^2}{\|X\|^2}$. Plugging this into the Bayes estimate, we recover the James-Stein estimator.

6.2 Extensions of James-Stein phenomenon (Lec 19)

Extension 1: The James-Stein phenomenon exists with any multivariate normal $X \sim \mathcal{N}(\mu, \Sigma)$, as long as the effective dimension is sufficiently large.

- The MLE is still X in this case.
- Let $\hat{\mu}_{JS} = \left(1 - \frac{(\tilde{p}-2)\sigma^2}{X^T \Sigma^{-1} X}\right) X$, where $\tilde{p} = \frac{\text{tr}(\Sigma)}{\lambda_{\max}(\Sigma)}$ is the **effective dimension**. If $\tilde{p} > 2$, then $R(\hat{\mu}_{JS}, \mu) < R(\hat{\mu}_{MLE}, \mu)$ for all $\mu \in \mathbb{R}^p$.

- **Linear regression context:** Consider the model $y = X\beta + \varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. If X has full column rank, then the James-Stein estimator $\hat{\beta}_{JS} = \left(1 - \frac{c}{\hat{\beta}_{MLE}^T X^T X \hat{\beta}_{MLE}}\right) X$ will dominate the MLE for MSE.

Extension 2: There is nothing special about shrinking to 0.

- Shrinking towards an arbitrary μ_0, i.e. $\hat{\mu}_{JS} = \mu_0 + \left(1 - \frac{(p-2)\sigma^2}{\|X - \mu_0\|^2}\right)(X - \mu_0)$ will also dominate the MLE.
- Instead of an arbitrary μ_0, we often use $\bar{X}$.

7 Model Selection (Lec 20)

Say we have the linear model $y = X\beta + z$, where $y \in \mathbb{R}^{n \times 1}$ observed, $X \in \mathbb{R}^{n \times p}$ is known, and $\beta \in \mathbb{R}^{p \times 1}$ is to be estimated. Assume $z_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$. We want to figure out which covariates to keep in our model.

- If y_i^* is the new observation at covariate value x_i, then the prediction error on observation i can be computed to be $\mathbb{E}[(y_i^* - \hat{y}_i)^2] = \mathbb{E}\left[\left(x_i^T\beta - x_i^T\hat{\beta}\right)^2\right] + \sigma^2$.
- Adding up across observations, we get **predictive risk** $PE = \sum_{i=1}^n \left[\mathbb{E}\left[\left(x_i^T\beta - x_i^T\hat{\beta}\right)^2\right] + \sigma^2\right] = \mathbb{E}\|X\beta - X\hat{\beta}\|^2 + n\sigma^2$.
- **Training error** is simply the residual sum of squares, i.e. $RSS = \|y - X\beta\|^2$. ("How well do I predict on the training set?")
- **Theorem:** $\mathbb{E}[RSS] < PE$. In fact, $\mathbb{E}[RSS] - PE = -2\sum_{i=1}^n \mathrm{Cov}(\hat{y}_i, y_i)$.

7.1 Linear estimation and C_p statistic (Lec 20)

Assume that our predictor is linear, i.e. $\hat{y} = My$ for some matrix M.

- Linear regression, ridge regression and smoothing splines are linear. LASSO is not linear.
- In this setting, $\mathbb{E}[RSS] - PE = -2\sigma^2 \mathrm{tr}(M)$, i.e. $PE = \mathbb{E}[RSS] + 2\sigma^2 \mathrm{tr}(M)$. This implies that $RSS + 2\sigma^2 \mathrm{tr}(M)$ is an unbiased estimate for prediction risk.
- When we select only a subset S of features, we get the OLS estimate $\hat{\beta}[S]$ and corresponding fitted values $\hat{y} = My$, where $M = X_S(X_S^T X_S)^{-1} X_S^T$. We can compute $\mathrm{tr}(M) = |S|$, thus giving the unbiased estimate for prediction risk: $RSS + 2|S|\sigma^2$.
- If we have an unbiased estimator for the variance, the C_p **statistic** is define as $C_p = RSS + 2\hat{\sigma}^2|S|$ (i.e. unbiased estimator for prediction risk.)

- An equivalent formulation of the C_p statistic is $\frac{RSS}{\hat{\sigma}^2} - n + 2|S|$.
- We can compute an unbiased estimate of prediction risk for ridge regression. See Lec 20 for details.

7.2 Model selection with C_p (Lec 21)

- **CAUTION:** If we do model selection with C_p: i.e. choose $S^* = \text{argmin}_S C_p(S)$, we run into trouble. Even though C_p is unbiased for the PE of each fixed model, $C_p(S^*)$ is NOT unbiased for $PE(S^*)$.
- Finding the best C_p model is equivalent to finding the estimate which solves $\text{argmin}_{\hat{\beta}}\|y-X\hat{\beta}\|^2+2\sigma^2\|\hat{\beta}\|_0$. (Generally computationally intractable.)
- Consider the special case where X is orthogonal: $X^TX = I$. In this case, solving for the best C_p model reduces to minimizing $\sum_{i=1}^{n}(y_i - \hat{\beta}_i)^2 + 2\sigma^2\|\hat{\beta}\|_0$. We can look at it coordinate by coordinate to obtain the solution
$$\hat{\beta}_i = \begin{cases} 0 & \text{if } |y_i| \leq \sqrt{2}\sigma, \\ y_i & \text{if } |y_i| > \sqrt{2}\sigma. \end{cases}$$
This is a **hard-thresholding rule**.

7.3 Model selection with the LASSO (Lec 22)

- Finding the best C_p model was equivalent to solving $\text{argmin}_{\hat{\beta}}\|y - X\hat{\beta}\|^2+\lambda^2\sigma^2\|\hat{\beta}\|_0$ with $\lambda^2 = 2$.
- We could be interested in other values of λ. The LASSO is a relaxation of this problem: $\text{argmin}_{\hat{\beta}}\|y - X\hat{\beta}\|^2+\lambda\sigma\|\hat{\beta}\|_1$.
- $\ell_0 - \ell_1$ **equivalence:** Under broad conditions, the minimizers of $\min\|\beta\|_{\ell_0}$ subject to $X\beta = y$ and $\min\|\beta\|_{\ell_1}$ subject to $X\beta = y$ are equal!
- **Solving the LASSO:** If we define $C = \{z \in \mathbb{R}^n : \|X^Tz\|_\infty \leq \lambda\}$ and let Π_C be the projection operator onto C, then we have $\hat{\mu} = y - \Pi_C(y)$ and $\hat{\beta} = X^\dagger\hat{\mu}$, where $X^\dagger$ is the pseudo-inverse of X.
- For the LASSO, we have $SURE = RSS + 2\sigma^2[n - \text{div }(\Pi_C(y))]$. Note that div $(\Pi_C(y))$ is simply the dimension of the affine space projected onto. Thus, $SURE(\lambda) = RSS(\lambda) + 2\sigma^2|\{j : \hat{\beta}_j(\lambda) \neq 0\}|$.

7.4 Oracle inequalities (Lec 23)

We have a model $y = X\beta + z$ and we want to choose the "best" submodel among $S \subseteq \{1, 2, \ldots, p\}$. For each subset S, let $\hat{\beta}[S]$ be the OLS regression coefficients and let $\hat{mu}[S] = X\hat{\beta}[S]$.

- Risk can be computed to be $R(\mu, \hat{\mu}[S]) = \|P_S\mu - \mu\|^2+|S|\sigma^2$, where P_S is the projection operator onto the subspace spanned by covariates in S. (Note that $\hat{\mu}[S] - P_S\mu$ is orthogonal to $P_S\mu - \mu$.)
- **Ideal risk** is defined as $R^I(\mu) = \min_S R(\mu, \hat{\mu}[S])$.
- If β is k-sparse (i.e. $\|\beta\|_0 \leq k$), then $R^I(\mu) \leq k\sigma^2$.

- In the case where $X = I$, i.e. $y \sim \mathcal{N}(\mu, \sigma^2 I)$, the risk of model S is $R(\mu, \hat{\mu}[S]) = \sum_{i \notin S} \mu_i^2 + |S|\sigma^2$.

  This is easy to minimize. From this we obtain $R^I(\mu) = \sum \min(\mu_i^2, \sigma^2)$, achieved when $\hat{\mu}_i^I = y_i$ if $|\mu_i| > \sigma$, 0 otherwise.

- **Can get estimator whose risk is close to ideal risk:** Suppose that we minimize

  $$\min \|y - X\hat{\beta}\|^2 + \lambda_p^2 \sigma^2 \|\hat{\beta}\|_0.$$

  If λ_p^2 is on the order of $2\log p$, then for all $\mu \in \mathbb{R}^n$,

  $$R(\mu, \hat{\mu}) \leq C_0 (2\log p)[\sigma^2 + R^I(\mu)],$$

  where C_0 is a constant that can be computed explicitly.

- **Theorem:** Suppose $Y \sim \mathcal{N}(\mu, \sigma^2 I)$. Let $\hat{\mu}$ be either a soft or hard thresholding estimator with $\lambda = \sigma\sqrt{2\log p}$. Then

  $$\mathbb{E}\|\mu - \hat{\mu}\|^2 \leq (2\log p + \delta) \left[\sigma^2 + \underbrace{\sum_i \min(\mu_i^2, \sigma^2)}_{R^I(\mu)} \right],$$

  with $\delta = 1$ for soft thresholding, $\delta = 1.2$ for hard thresholding. This inequality is not asymptotic, and it holds for any μ.

- **Risk inflation criterion:** Minimax result: Suppose $Y \sim \mathcal{N}(\mu, \sigma^2 I)$. For all estimators,

  $$\inf_{\hat{\mu}} \sup_{\mu} \frac{R(\mu, \hat{\mu})}{\sigma^2 + R^I(\mu)} \geq (2\log p)(1 + o_p(1)).$$

7.5 FDR thresholding (Lec 25)

As before, we have $y = X\mu + z$, $z \sim \mathcal{N}(0, \sigma^2 I)$. We wish to estimate μ, where $\mu \in \mathbb{R}^p$.

- **FDR hard thresholding estimator** is

  $$\hat{\mu}_{(i)} = \begin{cases} y_{(i)} & \text{if } |y_{(i)}| > t_{FDR}, \\ 0 & \text{otherwise.} \end{cases}$$

- The FDR hard thresholding estimator achieves near optimal guarantees: Under $X = I$, $\mu \in \ell_0(\varepsilon_n)$ with $\varepsilon_n \in \left[n^{-1}(\log n)^\delta, n^{-\delta}\right]$, the FDR estimator has the guarantee $\sup_{\mu \in \ell_0(\varepsilon)} \mathbb{E}\|\hat{\mu} - \mu\|^2 = \left(1 + \frac{(2q-1)_+}{1-q} + o_n(1)\right) R^*(\ell_0(\varepsilon))$.

- **SLOPE algorithm:** See Lec 25 for details.