Mathematical details for lecture series “Statistics for High-Dimensional Data”

1 Multiple Testing

1.1 Definitions

Definitions

$$
\begin{aligned}
H_0^i &: \text{ null-hypothesis for } i\text{th test} \\
\mathcal{H}_0 &: \{i : H_0^i \text{ is true}\} \\
\mathcal{H}_1 &: \{i : H_0^i \text{ is not true}\} \\
m &: |\mathcal{H}_0 \cup \mathcal{H}_1| \text{ number of tests} \\
\pi_0 &: |\mathcal{H}_0|/m \\
V &: \sum_{i \in \mathcal{H}_0} \mathbf{1}\{p_i < u\};\ \text{number of false positives given threshold } u \\
S &: \sum_{i \in \mathcal{H}_1} \mathbf{1}\{p_i < u\};\ \text{number of true positives given threshold } u \\
R &: \sum_{i} \mathbf{1}\{p_i < u\};\ \text{total number of rejections given threshold } u \\
\mathcal{R} &: \{i : p_i < u\};\ \text{the rejection set} \\
P_0 &: \text{ null probability measure.}
\end{aligned}
$$

1.2 Uniform distribution of p-values

Theorem
If a p-value, i.e. the probability that the test statistic $T(Y)$ exceeds observation $t = T(Y^*)$ given the null-hypothesis, $p = P_0(T(Y) > t)$, is computed from a continuous test statistic $T(Y)$, we have under the null-hypothesis

$$p(Y^*) \sim U[0,1] \text{ or equivalently } P_0(p(Y^*) \leq \alpha) = \alpha$$

.

Proof:

$$
\begin{aligned}
P_0(p(Y^*) \leq \alpha) &= P_0(P_0(T(Y) > T(Y^*)|T(Y^*)) \leq \alpha) = P_0(1 - F(T(Y^*)) \leq \alpha) \\
&= P_0(F(T(Y^*)) > 1 - \alpha) = P_0(T(Y^*) > F^{-1}(1-\alpha)) \\
&= 1 - P_0(T(Y^*) \leq F^{-1}(1-\alpha)) = 1 - F(F^{-1}(1-\alpha)) \\
&= 1 - (1-\alpha) = \alpha
\end{aligned}
\tag{1}
$$

1.3 Bonferroni and Holm

Proof Bonferroni: see Exercise

Proof Holm: Let $V^i = 0$ indicate the critical event that i is the smallest index for which null-hypothesis H_{0j} is true and $p_j = p_{(i)}$. Then $V^i = 0$ implies at most $m - (i-1)$ null-hypotheses.

$$\begin{aligned} P(V > 0) &= \sum_{i=1}^{m} P(V > 0|V^i = 0)P(V^i = 0) \leq \sum_{i=1}^{m}[\sum_{j=1}^{m} P(V_j = 1|V^i = 0)]P(V^i = 0) \\ &\leq \sum_{i=1}^{m}(m-(i-1))\alpha/((m-(i-1))P(V^i = 0) = \alpha, \end{aligned} \tag{2}$$

where we use the Bonferroni inequality for $P(V > 0|V^i = 0)$ and the fact that $P(V_j = 1|V^i = 0) = 0$ for all $(i-1)$ j's for which $p_j < p_{(i)}$.

1.4 Estimating π_0

A conservative estimate for the bFDR is obtained when replacing π_0 by 1. Quite a few methods exist for estimating π_0. A few simple estimators.

1. $\hat{\pi}_0$: fraction of non-rejected null-hypothesis when applying Benjamini-Hochberg.
2. $\hat{\pi}_0 = \min\left(1, \frac{\#\{p_i \in (0.25, 0.75)\}}{0.5m}\right)$. We count the number of times that p_i lies between the 25% and 75% quantile of the null (Uniform) and divide this by the number of times we expect this to happen if all null-hypotheses would hold: $0.5m$. Other pairs of quantiles $(q_{\lambda*100}, q_{(1-\lambda)*100})$ can be used; it may be tuned by applying a mean square error criterion.
3. Model the distribution of p-values as: $f(p) = \pi_0 + (1-\pi_0)h(p)$ and use the assumption $h(1) = 0$. Then, $\hat{\pi}_0 = \hat{f}(1)$. Parametric, semi-parametric and non-parametric methods may be used to model $f(p)$.

1.5 FDR estimation, bFDR

$$\begin{aligned} \text{bFDR} = P(i \in \mathcal{H}_0|p_i \leq t) &= \frac{P(i \in \mathcal{H}_0)P(p_i \leq t|i \in \mathcal{H}_0)}{P(p_i \leq t)} \\ &= \frac{\pi_0 F_0(t)}{F(t)}. \end{aligned} \tag{3}$$

Therefore, the following *estimate* is commonly used:

$$\text{bF}\hat{\text{D}}\text{R} = \frac{\hat{\pi}_0 m t}{\#\{p_i \leq t\}} = \frac{\hat{\pi}_0 F_0(t)}{\hat{F}(t)},$$

where $\hat{F}(t)$ denotes the empirical distribution of p-values.

bFDR is (sometimes) called "(Empirical) Bayesian FDR", because of the interpretation of $P(i \in \mathcal{H}_0 | p_i \le t)$ as a posterior probability.
Note: the Bayesian FDR, bFDR, is closely related another FDR concept pFDR $= E[V/R|R > 0]$. In fact, the same estimators are used for both.

Equivalence Theorem
The BH rule is equivalent to rejecting all H_0^i for which $p_i \le t$ and $t = \max_u\{\text{b}\hat{\text{FD}}\text{R}(u) \le \alpha\}$, when using $\pi_0 = 1$.

Proof
The above rejection rule is equivalent to using $p_i \le p_{(j)}$, where $p_{(j)}$ is the largest order statistic smaller than t. Then,

$$\text{b}\hat{\text{FD}}\text{R}(p_{(j)}) \le \alpha \Leftrightarrow \frac{P(p_i \le p_{(j)}|i \in \mathcal{H}_0)}{\hat{P}(p_i \le p_{(j)})} \le \alpha \Leftrightarrow \frac{p_{(j)}}{j/m} \le \alpha \Leftrightarrow p_{(j)} \le \alpha(j/m).$$

1.6 Local FDR, ℓfdr

bFDR can be written as $\pi_0 F_0(u)/F(u)$. A disadvantage of bFDR is that the significance for a p-value p_i (using bFDR cut-off $u = p_i$) depends on all values $\le p_i$. Local FDR, ℓfdr, is a more specific alternative:

$$\ell\text{fdr} = \pi_0 f_0(u)/f(u).$$

Here, usually $f_0 = 1$ is used (uniform distribution). Difficulty with ℓfdr is the estimation of $f(u)$, which is more difficult than estimating $F(u)$.

Note that $E_f[\ell\text{fdr}(u)|u \le v] = \text{b}\hat{\text{FD}}\text{R}(v)$, since

$$E_f[\pi_0 f_0(u)/f(u)|u \le v] = \pi_0\left(\int_{u \le v} f_0(u)/f(u) * f(u) du\right)/F(v) = \pi_0 F_0(v)/F(v).$$

Appendix (not compulsory): Proof of BH-rule controlling FDR

Let us write the FDR for given rejection set $\mathcal{R}$ as FDR($\mathcal{R}$), (NOTE: $|\mathcal{R}| = 0$ is not a concern, since this case does not contribute to the expectation) then

$$\begin{aligned}
\text{FDR}(\mathcal{R}) &= E\left[\frac{|\mathcal{R} \cap \mathcal{H}_0|}{|\mathcal{R}|}\right] \\
&= \sum_{i \in \mathcal{H}_0} E\left[\frac{\mathbf{1}\{i \in \mathcal{R}\}}{|\mathcal{R}|}\right] \\
&= \sum_{i \in \mathcal{H}_0} E\left[\frac{\mathbf{1}\{p_i \le \alpha|\mathcal{R}|/m\}}{|\mathcal{R}|}\right],
\end{aligned}$$

using that, $\mathcal{R}$ satisfies the self-consistency condition: $\mathcal{R} = \{i | p_i \leq \alpha |\mathcal{R}|/m\}$. To see this: $\mathcal{R} = \{i | p_i \leq p_{(J)}\}, J = \max(j | p_{(j)} \leq \alpha j/m)$, so $\mathcal{R} = \{i | p_i \leq \alpha J/m\}$. $|\mathcal{R}| = |\{i | p_i \leq p_{(J)}\}| = |\{i | p_i \leq \alpha J/m\}| = J$. Then, substitute J by $|\mathcal{R}|$: $\mathcal{R} = \{i | p_i \leq \alpha |\mathcal{R}|/m\}$.

The assertion $i \in \mathcal{R}$ is equivalent to $p_i \leq \alpha(|\mathcal{R}'_{-i}| + 1)/m$ and implies $|\mathcal{R}| = |\mathcal{R}'_{-i}| + 1$, where $\mathcal{R}'_{-i}$ is the rejection set $\{j \neq i | p_j \leq p_{(K)}\}, K = \max(k : p_{(k)} \leq \alpha(k+1)/m)$. Note that by definition, $\mathcal{R}'_{-i}$ only depends on the p-values of $\mathbf{p}_{-i} = (p_j, j \neq i)$. Therefore, the FDR can be rewritten as follows:

$$
\begin{aligned}
\mathrm{FDR}(\mathcal{R}) &= \sum_{i \in \mathcal{H}_0} E\left[\frac{\mathbf{1}\{p_i \leq \alpha[|\mathcal{R}'_{-i}| + 1]/m\}}{|\mathcal{R}'_{-i}| + 1}\right] \\
&= \sum_{i \in \mathcal{H}_0} E\left[E\left[\frac{\mathbf{1}\{p_i \leq \alpha[|\mathcal{R}'_{-i}| + 1]/m\}}{|\mathcal{R}'_{-i}| + 1} \;\middle|\; \mathbf{p}_{-i}\right]\right] \\
&= \sum_{i \in \mathcal{H}_0} E\left[\frac{E[\mathbf{1}\{p_i \leq \alpha[|\mathcal{R}'_{-i}| + 1]/m\}] | \mathbf{p}_{-i}]}{|\mathcal{R}'_{-i}| + 1}\right] \\
&= \sum_{i \in \mathcal{H}_0} E\left[\alpha/m\right] \\
&\leq \sum_{i \in \mathcal{H}_0 \cup \mathcal{H}_1} E\left[\alpha/m\right] \\
&= \frac{\alpha}{m} m, \\
&= \alpha
\end{aligned}
$$

where we used in the first inequality that the random variable p_i conditional to $\mathbf{p}_{-i}$ has the same distribution than its marginal (from the independence assumption).