High-dimensional data analysis, fall 2013

Yeast
understanding basic life functions
11,904 p-values
Blomberg et al. 2003, 2010

Arabidopsis Thaliana
association mapping
3,745 p-values
Zhao et al. 2007

fMRI brain scans
function of brain language network
appr. 3 mill. p-values
Taylor et al. 2006
Slides for B&vdG 11.1 – 11.5: p-values

Exercises: H:2
The standard "lowdimensional" way of selecting variables is to test if coefficients are significant. In Chapter 11 the authors try to extend this to highdimensional problems.

Setting: \( Y = X\beta + \epsilon + \) think of Lasso (but ideas more general)

\[
S_0 = \{j \in [1, \ldots, n]; \beta_j \neq 0\} \quad \text{true set of active variables}
\]

\( \hat{S} = \) estimated set of active variables

(\text{now} = \{j \in [1, \ldots, n]; p_j < \alpha\}, \text{with } p_j \text{ the } j-th \ p\text{-value}

\( V = \# \) not active covariates in \( \hat{S} = \# \) type 1 errors

\[
\text{FWER} := P(V > 0) \quad \text{FDR} := E\left(\frac{V}{\max(|\hat{S}|,1)}\right)
\]
A parenthesis: the Bonferroni method for error control

Type 1 error: reject a true null hypothesis (here = select a non-active variable)

Make $s$ tests, for each one compute a $p$-value, call this $p_j$

reject $j$-th null hypothesis if $p_j < \alpha/s$ ($\leftrightarrow p_js < \alpha$)

Then FEWR $= P(V > 0) = P($one or more false rejections$) \leq \alpha$.

PF: By Bonferroni’s inequality, with $P_0$ the distribution if all null hypotheses are true (using $p_i \sim U(0,1)$ under $P_0$)

\[ P(V > 0) \leq P_0(\{p_1 \leq \frac{\alpha}{s}\} \cup \cdots \cup \{p_s \leq \frac{\alpha}{s}\}) \]

\[ \leq P_0\left(p_1 \leq \frac{\alpha}{s}\right) + \cdots + P_0\left(p_s \leq \frac{\alpha}{s}\right) = s\alpha/s = \alpha. \]
Single sample splitting,

- Randomly split sample into two parts $I_1$ and $I_2$ of equal size $n/2$. Do variable selection on the first part, e.g. using one of the Lasso methods, to get $\hat{S}_0(I_1)$.

- Do a standard least square fit to the observations in $I_2$, using only the variables selected on the first part of the data, compute p-values $\tilde{P}_{raw,1}, \ldots, \tilde{P}_{raw,p}$ for the $p = |\hat{S}_0(I_1)|$ variables in $\hat{S}_0(I_1)$, using two-sided t-tests.

- Set $\tilde{P}_j = \begin{cases} \tilde{P}_{raw,j} & \text{if } j \text{ was selected} \\ 1 & \text{if } j \text{ was not selected} \end{cases}$ for $j = 1, \ldots, p$.

- Set $\tilde{P}_{corr,j} = \min\{\tilde{P}_j |\hat{S}_0|, 1\}$

- Make final selection: select the covariates for which $\tilde{P}_{corr,j} < \alpha$. 
Write $X(I_2)$ for the design matrix for analysis of the $I_2$ sample, and set

$$\hat{S}_{\text{single-split}}(\alpha) = \{j; \hat{P}_{\text{corr}, j} \leq \alpha\}$$

$$V_{\text{single-split}}(\alpha) = |\hat{S}_{\text{single-split}}(\alpha) \cap S_0^c|$$

**L. 11.1** Assume the noise $\epsilon$ is normally distributed, the design is fixed, and

$$\lim_{n \to \infty} P(S_0 \subset \hat{S}(I_1)) = 1$$

$$\lim_{n \to \infty} P(X(I_2)^tX(I_2) \text{ is invertible}) = 1.$$ 

Then the single-split method asymptotically controls FWER, i.e.

$$\limsup_{n \to \infty} P(V_{\text{single-split}}(\alpha) > 0) \leq \alpha.$$
**Pf.** On \( A_n = \{ S_0 \subset \hat{S}(I_1) \} \cup \{ X(I_2)^t X(I_2) \text{ is invertible} \} \) all the variables in \( \hat{S}(I_1)^c \) are inactive and hence can’t influence p-values. Further, the variables in \( I_2 \) are independent of the variables in \( I_1 \), the noise is Gaussian, and \( X(I_2)^t X(I_2) \) is invertible, and hence the t-test give exact p-values. The result then follows from the Bonferroni method, since \( P(A_n) \to 1 \) as \( n \to \infty \).

(It’s unclear to me if the assumption \( \lim_{n \to \infty} P(S_0 \subset \hat{S}(I_1)) = 1 \) really is needed for the result – however the method isn’t ”good” if this doesn’t hold)

There is one problem – but a big one – with this method: two persons which make the same analysis might get quite different results, because of the randomness in chosing \( I_1 \) and \( I_2 \).
The multi sample split method is defined as follows:
For $b = 1, \ldots, B$:

1. Randomly split the original data into two disjoint groups $I^{(b)}_1$ and $I^{[b]}_2$ of (almost) equal size.
2. Using only $I^{[b]}_1$, estimate the set of active predictors $\hat{S}^{[b]} = \hat{S}(I^{[b]}_1)$.
3. Using only $I^{[b]}_2$, compute the adjusted (non-aggregated) p-values as in (11.2), i.e.,

$$\tilde{P}^{[b]}_{\text{corr}, j} = \min(\tilde{P}^{[b]}_j \cdot |\hat{S}^{[b]}|, 1) \ (j = 1, \ldots, p)$$

where $\tilde{P}^{[b]}_j$ is based on the two-sided t-test, as in (11.1), based on $I^{[b]}_2$ and $\hat{S}^{[b]} = \hat{S}(I^{[b]}_1)$. 
The procedure described above leads to a total of $B$ p-values for each covariate $j = 1, \ldots, p$. For each $j = 1, \ldots, p$, the goal is to aggregate the p-values $\tilde{P}^{[b]}_{\text{corr},j}$ over the indices $b = 1, \ldots, B$. This can be done using quantiles. For $\gamma \in (0, 1)$ define

$$Q_j(\gamma) = \min \left\{ q_\gamma(\{ \tilde{P}^{[b]}_{\text{corr},j} / \gamma; b = 1, \ldots, B \}), 1 \right\},$$

(11.7)

where $q_\gamma(\cdot)$ is the (empirical) $\gamma$-quantile function.

A p-value for each variable $j = 1, \ldots, p$ is then given by $Q_j(\gamma)$, for any fixed $0 < \gamma < 1$. We will describe in Section 11.3.2 that this is an asymptotically correct p-value for controlling the familywise error rate.

A proper selection of $\gamma$ may be difficult. Error control is not guaranteed anymore if we search for the best value of $\gamma$. But we can use instead an adaptive version which selects a suitable value of the quantile based on the data. Let $\gamma_{\text{min}} \in (0, 1)$ be a lower bound for $\gamma$, typically 0.05, and define

Homework H.2: Explain (to me) why it makes sense to say that the $Q_j(\gamma)$ are “p-values”.
\[ P_j = \min \left\{ \left(1 - \log \gamma_{\text{min}}\right) \inf_{\gamma \in (\gamma_{\text{min}}, 1)} Q_j(\gamma), 1 \right\} \quad (j = 1, \ldots, p). \quad (11.8) \]

The extra correction factor \(1 - \log \gamma_{\text{min}}\) ensures that the familywise error rate remains controlled despite of the adaptive search for the best quantile, as described in Theorem 11.1 in Section 11.3.2. For the recommended choice of \(\gamma_{\text{min}} = 0.05\), this factor is upper bounded by 4; in fact, \(1 - \log(0.05) \approx 3.996\).
Fig. 11.2  Left: a histogram of adjusted p-values $\hat{P}^{[b]}_{\text{corr}, j}$ for a single variable in the motif regression example of Section 2.5.2 with $n = 287$ and $p = 195$. This is the same plot as in Figure 11.1. The single data split method picks randomly one of these p-values (a “p-value lottery”) and rejects $H_{0,j}$ if it is below $\alpha$. For the multi data split method, we reject $H_{0,j}$ if and only if the empirical distribution function of the adjusted p-values crosses from above the broken line (which is $f(p) = \max\{0.05, (3.996/\alpha)p\}$) for some value $p \in (0, 1)$ on the x-axis. This bound is shown as a broken line for $\alpha = 0.05$. For the given example, the bound is indeed exceeded and the variable is thus selected. The figure is taken from Meinshausen et al. (2009).
B&vdG 11.4: Multi sample splitting and FDR

\[ P_{(1)} \leq ... \leq P_{(p)} \text{ the ordered multi-split p-values} \]

\[ h = h(q) = \max\{i: P_{(i)} \leq iq\} \]

\[ \hat{S}_{\text{multisplit-FDR}}(\alpha) = \left\{ j; P_j \leq P_{(h(q(\alpha))} \right\}, \]

\[ q(\alpha) = \frac{\alpha}{\sum_{i=1}^{p} i^{-1}} \]

(If \( P_{(i)} > iq \) for all \( i \) then \( \hat{S}_{\text{multisplit-FDR}}(q) = \emptyset \))

Under stringent/strange conditions

\[ \lim_{n \to \infty} \sup \ FDR = E\left(Q(\alpha)\right) \leq \alpha \]

For \( Q(\alpha) = \frac{V_{\text{multisplit-FDR}}}{\max(R_{\text{multisplit-FDR}}(\alpha),1)} \)
\[ Y_i = \sum_{j=1}^{p} \beta_j X^j_i + \epsilon_i, \quad (i = 1, \ldots, n) \]

\[ \epsilon_1, \ldots, \epsilon_n \text{ i.i.d. } \sim N(0, \sigma) \]

\[ X_1, \ldots, X_n \text{ i.i.d. } N(0, \Sigma), \quad \Sigma = \begin{pmatrix} 1 & \cdots & \rho \\ \vdots & \ddots & \vdots \\ \rho & \cdots & 1 \end{pmatrix} \]

\[ s_0 = |S_0| \text{ non-zero } \beta_j, \text{ either } = 1, \ldots, 1 \text{ or } = 1, 2, \ldots, s_0 \]

\[ B = \text{ the number of splits } = 100 \]

Initial selection:

\[ \hat{S}_{\text{fixed}} \text{ the } \lfloor n/6 \rfloor \text{ “best” variables in Lasso regularization path} \]

\[ \hat{S}_{\text{CV}} \text{ 10-fold crossvalidated Lasso} \]

\[ \hat{S}_{\text{adapt}} \text{ the adaptive Lasso starting with } \hat{S}_{\text{CV}} \]
S the single-split method, M the multi-split method, FEWR asymptotically controlled at 0.05 for both methods
<table>
<thead>
<tr>
<th>Uniform Sampling</th>
<th>$S_0$</th>
<th>SNR</th>
<th>( E(\text{ True Positives} )</th>
<th>( E(\text{ False Positives} )</th>
<th>( P(\text{ False Positives &gt; 0} )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Multi Split</td>
<td>Adaptive Lasso</td>
<td>Multi Split</td>
</tr>
<tr>
<td>NO</td>
<td>10</td>
<td>0.25</td>
<td>0.00</td>
<td>2.30</td>
<td>0</td>
</tr>
<tr>
<td>NO</td>
<td>10</td>
<td>1</td>
<td>0.58</td>
<td>6.32</td>
<td>0</td>
</tr>
<tr>
<td>NO</td>
<td>10</td>
<td>4</td>
<td>4.14</td>
<td>8.30</td>
<td>0</td>
</tr>
<tr>
<td>NO</td>
<td>10</td>
<td>16</td>
<td>7.20</td>
<td>9.42</td>
<td>0.02</td>
</tr>
<tr>
<td>YES</td>
<td>10</td>
<td>0.25</td>
<td>0.02</td>
<td>2.52</td>
<td>0</td>
</tr>
<tr>
<td>YES</td>
<td>10</td>
<td>1</td>
<td>0.10</td>
<td>7.46</td>
<td>0.02</td>
</tr>
<tr>
<td>YES</td>
<td>10</td>
<td>4</td>
<td>2.14</td>
<td>9.96</td>
<td>0</td>
</tr>
<tr>
<td>YES</td>
<td>10</td>
<td>16</td>
<td>9.92</td>
<td>10.00</td>
<td>0.04</td>
</tr>
<tr>
<td>NO</td>
<td>5</td>
<td>0.25</td>
<td>0.06</td>
<td>1.94</td>
<td>0</td>
</tr>
<tr>
<td>NO</td>
<td>5</td>
<td>1</td>
<td>1.50</td>
<td>3.86</td>
<td>0.02</td>
</tr>
<tr>
<td>NO</td>
<td>5</td>
<td>4</td>
<td>3.52</td>
<td>4.58</td>
<td>0.02</td>
</tr>
<tr>
<td>NO</td>
<td>5</td>
<td>16</td>
<td>4.40</td>
<td>4.98</td>
<td>0</td>
</tr>
<tr>
<td>YES</td>
<td>5</td>
<td>0.25</td>
<td>0.02</td>
<td>2.22</td>
<td>0</td>
</tr>
<tr>
<td>YES</td>
<td>5</td>
<td>1</td>
<td>0.82</td>
<td>4.64</td>
<td>0.02</td>
</tr>
<tr>
<td>YES</td>
<td>5</td>
<td>4</td>
<td>4.90</td>
<td>5.00</td>
<td>0</td>
</tr>
<tr>
<td>YES</td>
<td>5</td>
<td>16</td>
<td>5.00</td>
<td>5.00</td>
<td>0</td>
</tr>
</tbody>
</table>

\( N=100, \ p=200 \)
Light bars: standard FDR control (only possible for $p < n$)
Dark bars: Multisplit method

$$\text{FDR} = E\left(\frac{V}{\max(|\hat{S}|,1)}\right) \text{ controlled at level } \alpha = 0.05$$