hdi: High-dimensional inference

Ruben Dezeure
ETH Zürich

based on joint work with

Peter Bühlmann
Sara van de Geer
Lukas Meier
Nicolai Meinshausen
Outline of the talk

1. Motivating example: biomarker discovery
2. Measures of performance
3. Current state of the art
Motivating example

Biomarker discovery is a medical term describing the process by which biomarkers are discovered. (Wikipedia)

The term biomarker ... refers to ... objective indications of medical state observed from outside the patient which can be measured accurately and reproducibly. ... in contrast to medical symptoms, which are limited to those indications of health or illness perceived by patients themselves. (Strimbu, Tavel (2010))
Good biomarkers

Three important technical attributes:

1. Must be present in peripheral body tissue and/or fluid
2. It must be easy to detect or quantify in the lab in procedures that are both affordable and robust
3. Its appearance must be associated as specifically as possible with damage of a particular tissue

(Editorial, Nature biotechnology 28, 431 (2010))
What problem are we dealing with here?

- Prediction?
- Variable selection?
- Hypothesis testing

\[ Y = X\beta^0 + \varepsilon, \quad X \text{ is } n \times p \text{ design} \]

\[ H_{0,j} : \beta_j^0 = 0 \]

\[ H_{A,j} : \beta_j^0 \neq 0 \]
What problem are we dealing with here?

- Prediction?
- Variable selection?
- Hypothesis testing

\[ Y = X\beta^0 + \varepsilon, \quad X \text{ is } n \times p \text{ design} \]

\[ H_{0,j} : \beta_j^0 = 0 \]

\[ H_{A,j} : \beta_j^0 \neq 0 \]
What problem are we dealing with here?

- Prediction?
- Variable selection?
- Hypothesis testing

\[ Y = X\beta^0 + \varepsilon, \quad X \text{ is } n \times p \text{ design} \]

\[ H_{0,j} : \beta_j^0 = 0 \]

\[ H_{A,j} : \beta_j^0 \neq 0 \]
Isn’t this just Stats 101?

Although solved in the low-dimensional case
($\#\text{features} < \#\text{samples}$)

It was, until recently, very much an open problem for the high-dimensional case!

Screening many possible biomarkers at the same time
→ not trivial
Isn’t this just Stats 101?

Although solved in the low-dimensional case (#features < #samples)

It was, until recently, very much an open problem for the high-dimensional case!

Screening many possible biomarkers at the same time → not trivial
Recent proposals

- Multi sample splitting (Meinshausen, Meier and Bühlmann (2009))
- Ridge projection method (Bühlmann (2013))
- De-sparsified Lasso (Zhang and Zhang (2013))
  (van de Geer, Bühlmann, Ritov and Dezeure (2014))
- (Javanmard and Montanari (2013))
- Covtest (Lockhart, Taylor, Tibshirani and Tibshirani (2014))

Varying in theoretical results and computational effort
Theoretical perspective

We desire

\[ \sqrt{n}(\hat{b}_j - \beta^0_j) \overset{\text{d}}{\rightarrow} \mathcal{N}(0, \omega^2_j), \]

with an explicit expression for \( \omega^2_j \).

Assumptions:

- Sparsity: \( s_0 = o \left( \frac{\sqrt{n}}{\log(p)} \right) \) vs.
- \( o \left( \frac{n}{\log(p)} \right) \)
- Design \( X \) not too ill-posed, **Restricted eigenvalue assumption**
- \( X \sim \mathcal{N}(0, \Sigma) \)
- \( \Sigma^{-1} \) row-sparse: \( \max_j s_j = o \left( \frac{n}{\log(p)} \right) \)
Theoretical perspective

We desire

\[ \sqrt{n}(\hat{b}_j - \beta_j^0) \Rightarrow \mathcal{N}(0, \omega_j^2), \]

with an explicit expression for \( \omega_j^2 \).

Assumptions:

- **Sparsity**: \( s_0 = o\left(\frac{\sqrt{n}}{\log(p)}\right) \) vs \( o\left(\frac{n}{\log(p)}\right) \)
- **Design** \( X \) not too ill-posed, *Restricted eigenvalue assumption*
- \( X \sim \mathcal{N}(0, \Sigma) \)
- \( \Sigma^{-1} \) row-sparse: \( \max_j s_j = o\left(\frac{n}{\log(p)}\right) \)
# Performance measures

Repeated experiments

<table>
<thead>
<tr>
<th>True biomarkers</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

- **Power** = TPR
- **Error rate** = fraction of experiments with at least 1 FP ≠ FDR ≤ α = 0.05
### Performance measures

**Repeated experiments**

<table>
<thead>
<tr>
<th></th>
<th>True biomarkers</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>✓</td>
<td>✗</td>
</tr>
<tr>
<td>100</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

- **Power** (TPR) = \( \frac{TP}{TP+FN} \)
- **Error rate** (FDR) = \( \frac{FP}{FP+TP} \)
- \( \alpha \) ≤ 0.05
**Performance measures**

**Repeated experiments**

<table>
<thead>
<tr>
<th>True biomarkers</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>99</td>
<td></td>
</tr>
<tr>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

- True biomarkers
- Others

```
1  ✔   ☐   ☐
2  ☐   ☐   ☐
:  :   :   :
99 ✔   ☐   ✔
100 ✔   ☐   ☐
```

**Power = TPR**

Error rate = fraction of experiments with at least 1 FP \( \leq \alpha = 0.05 \)
Performance measures

Repeated experiments

<table>
<thead>
<tr>
<th>True biomarkers</th>
<th>Others</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>☑</td>
</tr>
<tr>
<td>2</td>
<td>☐</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>99</td>
<td>☑</td>
</tr>
<tr>
<td>100</td>
<td>☑</td>
</tr>
</tbody>
</table>

Power = TPR

Error rate = fraction of experiments with at least 1 FP $\neq$ FDR

$\leq \alpha = 0.05$
Comparison plots
(Dezeure, R., Bühlmann, P., Meier, L., Meinshausen, N. (2014))

$s_0 = 3$

$s_0 = 15$

Source code available as R package *hdi*
Further work

Still quite a few assumptions being made

- Existing methods should be improved to handle some assumption violations

Power will disappear for really high dimensions due to error rate control

- Group testing as alternative
- Selective inference
Further work

Still quite a few assumptions being made

- Existing methods should be improved to handle some assumption violations

Power will disappear for really high dimensions due to error rate control

- Group testing as alternative
- Selective inference
Take home messages

- Prominent problem in biomarker discovery
- Performance can be quite good
- Work to be done on required assumptions
Thank you!
References:

References part 2:


