Ranking and Selection in High-Dimensional Inference

Michael Newton       Nick Henderson

in memory of John Klein
• large-scale
• not sparse
• ranking/sorting/prioritizing
• variance artifacts
• agreement
• empirical Bayes
• r-values
Example Type 2 Diabetes (T2D) GWAS (Morris et al. 2012, Nat Gen)

• case/control (22,669 / 58,119)

• lots of T2D associated loci, but of small effect (3371 SNPs shown)

• ?how to rank order?

arXiv:1312.5776
Example

Type 2 Diabetes (T2D) GWAS (Morris et al. 2012, Nat Gen)

- case/control (22,669 / 58,119)
- lots of T2D associated loci, but of small effect (3371 SNPs shown)
- how to rank order?

\[ \log \frac{\text{odds}(T2D|A)}{\text{odds}(T2D|A^c)} \]

arXiv:1312.5776
Example

Type 2 Diabetes (T2D) GWAS (Morris et al. 2012, Nat Gen)

- case/control (22,669 / 58,119)
- lots of T2D associated loci, but of small effect (3371 SNPs shown)
- how to rank order?
Example  

**gene-set enrichment, RNAi** *(Hao et al. 2013, PLoS Comp Bio)*

- 984 human genes linked to influenza-virus replication
- Functional content measured against Gene Ontology (5719 sets)
- ?how to rank order?

*arXiv:1312.5776*
• 461 NBA players (2013–2014)
• free throw percentage
• ?how to rank order?
simulation

\[
\begin{align*}
\text{units} & \quad i = 1, 2, \ldots, B \\
\text{signals} & \quad \theta_i \\
\text{f} & \quad N(0, 1) \\
\text{noise levels} & \quad \sigma_i^2 \\
\text{measured signals} & \quad X_i \\
\text{g} & \quad \text{Gam}(a, b)
\end{align*}
\]
data pairs: \((X_1, \sigma_1^2), (X_2, \sigma_2^2), \cdots, (X_B, \sigma_B^2)\)

ranking statistic: \(R_1, R_2, \ldots, R_B\)

aim: highly rank units with largest signals

arXiv:1312.5776
lead units by \textbf{p-value} are enriched for those with small variance

\[ p(\sigma_i \mid \text{p-value}_i \leq p_{0.1}) \]
lead units by p-value are enriched for those with small variance

\[ p(\sigma_i | \text{p-value}_i \leq p_{0.1}) \]

same for q-value!
other approaches

rank by “local” maximum likelihood estimate

• estimated log odds ratio
• proportion of gene set on gene list
• free throw percentage
lead units by MLE are enriched for those with large variance

\[ p(\sigma_i \mid X_i \geq x_{0.1}) \]

\[ p(\sigma_i) \]

local MLE

arXiv:1312.5776
lead units by posterior mean are enriched for those with small variance

\[ p \{ \sigma_i \mid E(\theta_i \mid X_i, \sigma_i) \geq e_{0.1} \} \]
Problems and solutions


We’ve found a generic empirical bayes ranking/selection method
• a ranking method corresponds to a family of threshold functions:

\[ T = \{ t_\alpha : \alpha \in (0, 1) \} \]
• a ranking method corresponds to a family of threshold functions:

\[ \mathcal{T} = \{ t_\alpha : \alpha \in (0, 1) \} \]

• each one is a function

[Graph showing a relationship between \( X \) and \( t_\alpha(\sigma^2) \)]
• a ranking method corresponds to a family of threshold functions:

\[ \mathcal{T} = \{t_\alpha : \alpha \in (0, 1)\} \]

• each one is a function

• unit i ranked in top \( \alpha \) iff \( X_i \geq t_\alpha(\sigma_i^2) \)
a ranking method corresponds to a family of threshold functions:

\[ \mathcal{T} = \{ t_\alpha : \alpha \in (0, 1) \} \]

- each one is a function

- unit i ranked in top \( \alpha \) iff \( X_i \geq t_\alpha(\sigma_i^2) \)

- size constraint: \( P \{ X_i \geq t_\alpha(\sigma_i^2) \} = \alpha \) (marginal!!)

arXiv:1312.5776
T2D example

rank by sweeping through the family

\[ X \]

\[ \sigma^2 \]

a. MLE

b. p-value

arXiv:1312.5776
\( f = N(0,1) \)

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Ranking Variable</th>
<th>Threshold Function ( t_\alpha(\sigma^2) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>( X_i )</td>
<td>( u_\alpha )</td>
</tr>
<tr>
<td>PV ( H_0 : \theta_i = 0 )</td>
<td>( X_i/\sigma_i )</td>
<td>( u_\alpha \sigma )</td>
</tr>
<tr>
<td>PV ( H_0 : \theta_i = c )</td>
<td>( (X_i - c)/\sigma_i )</td>
<td>( c + u_\alpha \sigma )</td>
</tr>
<tr>
<td>PM</td>
<td>( X_i/(\sigma_i^2 + 1) )</td>
<td>( u_\alpha(\sigma^2 + 1) )</td>
</tr>
<tr>
<td>PER</td>
<td>( P(\theta_i \leq \theta</td>
<td>X_i, \sigma_i^2) )</td>
</tr>
<tr>
<td>BF</td>
<td>( 1(X_i &gt; 0) \frac{P(X_i</td>
<td>\sigma_i^2, \theta_i \neq 0)}{P(X_i</td>
</tr>
</tbody>
</table>


arXiv:1312.5776
Even though the unit-level parameters are unobserved, their distribution may be well estimated.

Kiefer Wolfowitz, 1956

We can estimate $\theta_\alpha$ such that:

$$P(\theta_i \geq \theta_\alpha) = \int_{\theta_\alpha}^{\infty} f(\theta) \, d\theta = \alpha$$
\[
P \left\{ X_i \geq t_\alpha(\sigma_i^2), \theta_i \geq \theta_\alpha \right\}
\]

reported in top fraction

truly in top fraction
Theorem 1: Under certain smoothness conditions, 

$$\{t^*_\alpha\} \text{ is optimal if for all } \sigma^2$$

$$P \left\{ \theta_i \geq \theta_\alpha \mid X_i = t^*_\alpha(\sigma^2), \sigma^2_i = \sigma^2 \right\} = c_\alpha$$
### Table 1.

<table>
<thead>
<tr>
<th>criteria</th>
<th>ranking variable</th>
<th>threshold function $t_\alpha(\sigma^2)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLE</td>
<td>$X_i$</td>
<td>$u_\alpha$</td>
</tr>
<tr>
<td>PV $H_0: \theta_i = 0$</td>
<td>$X_i/\sigma_i$</td>
<td>$u_\alpha\sigma$</td>
</tr>
<tr>
<td>PV $H_0: \theta_i = c$</td>
<td>$(X_i - c)/\sigma_i$</td>
<td>$c + u_\alpha\sigma$</td>
</tr>
<tr>
<td>PM</td>
<td>$X_i/(\sigma_i^2 + 1)$</td>
<td>$u_\alpha(\sigma^2 + 1)$</td>
</tr>
<tr>
<td>PER</td>
<td>$P(\theta_i \leq \theta</td>
<td>X_i, \sigma_i^2)$</td>
</tr>
<tr>
<td>BF</td>
<td>$1(X_i &gt; 0)\frac{P(X_i</td>
<td>\sigma_i^2, \theta_i \neq 0)}{P(X_i</td>
</tr>
</tbody>
</table>

Maximal agreement can solve directly if $f = N(0,1)$

---


arXiv:1312.5776
maximal agreement

\[
X
\]

\[
\sigma^2
\]

d. maximal agreement

arXiv:1312.5776
maximal agreement

pretty close, considering that we’re targeting agreement not the artifact

\[
p \{ \sigma_i \mid X_i \geq t_{0.1}^* (\sigma_i^2) \}
\]
Theorem 2: Under certain conditions, the optimal family $\{t^*_\alpha\}$ satisfies:

$$t^*_\alpha(\sigma^2) = \inf \{x : V_\alpha(x, \sigma^2) \geq \lambda_\alpha\}$$

where,

$$P \left\{ V_\alpha(X_i, \sigma^2_i) \geq \lambda_\alpha \right\} = \alpha$$

$$V_\alpha(X_i, \sigma^2_i) = P\left( \theta_i \geq \theta_\alpha | X_i, \sigma^2_i \right)$$
from thresholds back to ranking variables
\[ r_i(X_i, \sigma_i^2) = \inf \{ \alpha : X_i \geq t_\alpha(\sigma_i^2) \} \]
\[ r_i(X_i, \sigma_i^2) = \inf \{ \alpha : X_i \geq t_\alpha(\sigma_i^2) \} \]

\[ r_i(X_i, \sigma_i^2) = \inf \{ \alpha : V_\alpha(X_i, \sigma_i^2) \geq \lambda_\alpha \} . \]
smallest $\alpha$ such that unit $i$ in top $\alpha$
when ranking by: $V_{\alpha}(X_i, \sigma_i^2) = P(\theta_i \geq \theta_\alpha | X_i, \sigma_i^2)$
The r-value concept makes sense in various elaborations of the theorems from Section 2.1. We retain univariate parameters of interest varying according to a distribution $F_i$. Further, the top fraction of units by r-value has higher overlap with the true top list, and this is closely related to a Bayes optimal ranking under a certain loss function.

Let $\lambda_\alpha = H_{\alpha-1}(1-\alpha)$. Then by analogy to (8), the r-value is defined:

$$r(D_i) = \inf \{ \alpha : V_\alpha(D_i) \geq \lambda_\alpha \}.$$
how to calculate r-values
Binomial likelihood
e.g., Beta prior/posteriors
Binomial likelihood
e.g., Beta prior/posteriors
P( \theta_i \geq \theta_\alpha | D_i )

two examples

empirical quantile

\lambda

r-value

D_{Ray.\,Allen} = 105/116

D_{LeBron.\,James} = 439/585

\alpha
## Top 10

<table>
<thead>
<tr>
<th>Player</th>
<th>Free throw %</th>
<th>r-value</th>
<th>post. mean</th>
<th>qual. rank</th>
<th>FTP rank</th>
<th>PM rank</th>
<th>RV rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brian Roberts</td>
<td>125/133</td>
<td>94.0</td>
<td>0.002</td>
<td>91.3</td>
<td>1</td>
<td>17</td>
<td>1</td>
</tr>
<tr>
<td>Ryan Anderson</td>
<td>59/62</td>
<td>95.2</td>
<td>0.003</td>
<td>89.8</td>
<td>15</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Danny Granger</td>
<td>63/67</td>
<td>94.0</td>
<td>0.005</td>
<td>89.3</td>
<td>16</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Kyle Korver</td>
<td>87/94</td>
<td>92.6</td>
<td>0.008</td>
<td>89.2</td>
<td>19</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Mike Harris</td>
<td>26/27</td>
<td>96.3</td>
<td>0.010</td>
<td>86.6</td>
<td>14</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>J.J. Redick</td>
<td>97/106</td>
<td>91.5</td>
<td>0.011</td>
<td>88.6</td>
<td>22</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Ray Allen</td>
<td>105/116</td>
<td>90.5</td>
<td>0.016</td>
<td>88.0</td>
<td>25</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>Mike Muscala</td>
<td>14/14</td>
<td>100.0</td>
<td>0.017</td>
<td>84.4</td>
<td>7</td>
<td>34</td>
<td>8</td>
</tr>
<tr>
<td>Dirk Nowitzki</td>
<td>338/376</td>
<td>89.9</td>
<td>0.018</td>
<td>89.1</td>
<td>30</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>Trey Burke</td>
<td>102/113</td>
<td>90.3</td>
<td>0.018</td>
<td>87.7</td>
<td>28</td>
<td>9</td>
<td>10</td>
</tr>
</tbody>
</table>
Predictive accuracy

\[ E \left[ \text{similarity}_t \left( \text{Ranks}(\theta), \text{Ranks.hat}[\text{midseason}] \right) \mid \text{complete season} \right] \]

- **r-value**
- **posterior mean**
- **MLE**

\[ t = \text{rank from top} \]
Fig. 3. Threshold functions, T2D example, data and axes as in Fig 1: Calculations use an inverse-gamma model for $\sigma^2$. Forty two threshold functions are shown, ranging in $\alpha$ values from a small positive value (red) just including the first data point up to $\alpha=0.1$ (blue). (Most data points are truncated by the plot, as in Fig 1; also, the grid is uniform on the scale of $\log_2(\alpha)$. Units associated with a smaller $\alpha$ (i.e., more red) are ranked more highly by the given ranking method. Two units landing on the same curve would be ranked in the same position.

a. MLE  b. p-value  c. posterior mean  d. maximal agreement
\[ \sigma^2 \sim \text{Exp}(1) \]
multi-loss

\[ L_\alpha(a, \theta_i) = 1 - 1(a \leq \alpha, \theta_i \geq \theta_\alpha) \]

risk_\alpha = 1 - P \{\delta(D_i) \leq \alpha, \theta \geq \theta_\alpha\},

marginally constrained

\[ \text{risk}_\alpha + \gamma_\alpha P \{\delta(D_i) \leq \alpha\} \]
summary

- ranking to maximize agreement
- large-scale, non-sparse settings
- empirical Bayes inference
- R-package: rvalues