Making the cut: improved ranking and selection for large-scale inference

Nicholas C. Henderson and Michael A. Newton

University of Wisconsin, Madison, USA
scale

error-free measurement
scale

measurement

estimate +/- 2 SE
multivariate
multivariate

what we want
multivariate

what we get
large scale

- increasing estimate

- regression effect
- variance effect
large scale

-4 -2 0 2 4 6 8

increasing estimate/SE
• large-scale
• not sparse
• ranking/sorting/prioritizing
• variance artifacts
• agreement
• empirical Bayes
• r-values
• case/control
  \(22,669 / 58,119\)

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

• ?how to rank order?

\[ \log \frac{\text{odds}(\text{T2D}|A)}{\text{odds}(\text{T2D} | A^c)} \]
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?
- 461 NBA players (2013-2014)
- free throw percentage
- ?how to rank order?
signals

\( \theta_i \)

\( f = N(0,1) \)

noise levels

\( \sigma_i^2 \)

\( g = \text{Gam}(a,b) \)

measured signals

\( X_i \)

\( N(\theta_i, \sigma_i^2) \)

units \( i = 1, 2, \ldots, B \)
simulation

data pairs: \((X_1, \sigma_1^2), (X_2, \sigma_2^2), \ldots, (X_B, \sigma_B^2)\)

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

aim: highly rank units with largest signals
lead units by p-value are enriched for those with small variance

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

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

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

\[ p(\sigma_i) \]
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:

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

• each one is a function

\[ t_\alpha(\sigma^2) \]

• unit i ranked in top \( \alpha \) if \( X_i \geq t_\alpha(\sigma_i^2) \)

• size constraint: \( P \left\{ X_i \geq t_\alpha(\sigma_i^2) \right\} = \alpha \) (marginal!!)
T2D example

rank by sweeping through the family

-X-

\( X \)

\( \sigma^2 \)

a. MLE

b. p-value
Table 1. Threshold functions associated with various ranking criteria, normal/normal model

<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>
Aside on PER

\[
\text{rank}(\theta_i) = \sum_{j=1}^{n} 1[\theta_i \leq \theta_j]
\]

\[
E \left[ \text{rank}(\theta_i)/n | \text{data} \right] = \frac{1}{n} \sum_{j=1}^{n} P[\theta_i \leq \theta_j | \text{data}]
\]

\[
P(\theta_i \leq \theta | X_i, \sigma^2_i)
\]

\[
u_\alpha \sqrt{(\sigma^2 + 1)(2\sigma^2 + 1)}
\]

**PER**

- posterior
- expected rank

large-scale limit

threshold function
\[ P \left\{ X_i \geq t_\alpha(\sigma_i^2), \theta_i \geq \theta_\alpha \right\} \]

reported in top fraction

\[ P(\theta_i \geq \theta_\alpha) = \int_{\theta_\alpha}^{\infty} f(\theta) \, d\theta = \alpha \]
Theorem 1: In the measurement model,

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

\[ P \{ \theta_i \geq \theta_\alpha | X_i = t^*_\alpha(\sigma^2), \sigma^2_i = \sigma^2 \} = c_\alpha \]

Measurement model: joint density for \( (X_i, \theta_i, \sigma^2_i) \)

\[ \theta_i = E(X_i|\theta_i, \sigma^2_i) \]
\[ \sigma^2_i = var(X_i|\theta_i, \sigma^2_i) \]

\(^\dagger\text{among continuously differentiable thresholds}\)
maximal
agreement

<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>$1(X_i &gt; 0) \frac{P(X_i</td>
<td>\sigma_i^2, \theta_i \neq 0)}{P(X_i</td>
</tr>
<tr>
<td>max agreement</td>
<td>r-value</td>
<td>$\theta_\alpha(\sigma^2 + 1) - u_\alpha \sqrt{\sigma^2(\sigma^2 + 1)}$</td>
</tr>
</tbody>
</table>

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

d. maximal agreement
maximal agreement

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

\[ p \left\{ \sigma_i \mid X_i \geq t^*_{0.1}(\sigma_i^2) \right\} \]
local tail probability

\[ V_\alpha(X_i, \sigma_i^2) = P(\theta_i \geq \theta_\alpha | X_i, \sigma_i^2) \]

Theorem 2: If \( V_\alpha(x, \sigma^2) \) is non-decreasing and right continuous in \( x \) for every \( \alpha, \sigma^2 \), and

\[ P \{ V_\alpha(X_i, \sigma_i^2) \geq \lambda_\alpha \} = \alpha \]

then,

\[ t^*_\alpha(\sigma^2) = \inf \{ x : V_\alpha(x, \sigma^2) \geq \lambda_\alpha \} \]
Theorem 3: If $V_\alpha(x, \sigma^2)$ is strictly increasing and differentiable in $\alpha$ for each $(x, \sigma^2)$, and if $\lambda_\alpha$ is differentiable, then

a. $t^*_\alpha(\sigma^2) = V_{\alpha}^{-1}(\lambda_\alpha, \sigma^2)$

b. $\frac{\partial}{\partial \alpha} t^*_\alpha(\sigma^2) < 0 \quad \forall \sigma^2 \Rightarrow \text{no crossing}$

Corollary: No crossing in normal/normal
from thresholds back to ranking variables
The proposed ranking procedure is a kind of Bayes rule for a percentile, with the rule for loss which is one minus the agreement (2). In the absence of other considerations, the additional structure that it shares with the Bayes rule then minimizes the modified objective function:

\[ r_i(X_i, \sigma_i^2) = \inf \{ \alpha : X_i \geq t_{\alpha}(\sigma_i^2) \} \]

r-value

\[ r_i(X_i, \sigma_i^2) = \inf \{ \alpha : V_{\alpha}(X_i, \sigma_i^2) \geq \lambda_{\alpha} \} . \]
smallest $\alpha$ such that unit $i$ in top $\alpha$
when ranking by: $V_{\alpha}(X_i, \sigma_i^2) = P(\theta_i \geq \theta_\alpha | X_i, \sigma_i^2)$
Also, the conditional tail probability has been used for ranking (like for all the families shown in Figure 3 and Table 1, the optimal threshold does not touch agreement at all list sizes, the proposed method does not require a measurement model. Instead, we retain univariate parameters of interest from Section 2.1. We retain univariate parameters of interest to generate its ranking. Conceptually, we imagine ranking the units by their r-value (not just a pre-specified one); then the r-value for unit $X_i$ is

$$r(D_i) = \inf \{\alpha : V_\alpha(D_i) \geq \lambda_\alpha\}.$$ 

Like for all the families shown in Figure 3 and Table 1, the optimal threshold establishes the role of this conditional tail probability in maximizing an agreement at all list sizes, the proposed method does not require a pre-specified threshold function. For the thresholds (7) that are optimal for agreement between different ranking procedures, it also happens to land at position $H_\sigma \in \{x, \lambda\}$ separately for all possible values of $\sigma$. Induced by the marginal distribution of $(\theta, \lambda, \sigma)$, the critical distance among units,

$$V_\alpha(D_i) = P(\theta_i \geq \theta_\alpha | D_i)$$

$$\lambda_\alpha = H^{-1}_\alpha(1 - \alpha).$$

We also retain the assumption of mutual independence in Theorem 3. The critical distance is $\sigma_i$ for every unit, and it is continuous in $\sigma_i$. For the thresholds (7) that are optimal for agreement between different ranking procedures, it also happens to land at position $H_\sigma \in \{x, \lambda\}$ separately for all possible values of $\sigma$. Induced by the marginal distribution of $(\theta, \lambda, \sigma)$, the critical distance among units,

$$V_\alpha(D_i) = P(\theta_i \geq \theta_\alpha | D_i)$$

$$\lambda_\alpha = H^{-1}_\alpha(1 - \alpha).$$

We also retain the assumption of mutual independence in Theorem 3. The critical distance is $\sigma_i$ for every unit, and it is continuous in $\sigma_i$. For the thresholds (7) that are optimal for agreement between different ranking procedures, it also happens to land at position $H_\sigma \in \{x, \lambda\}$ separately for all possible values of $\sigma$. Induced by the marginal distribution of $(\theta, \lambda, \sigma)$, the critical distance among units,

$$V_\alpha(D_i) = P(\theta_i \geq \theta_\alpha | D_i)$$

$$\lambda_\alpha = H^{-1}_\alpha(1 - \alpha).$$

We also retain the assumption of mutual independence in Theorem 3. The critical distance is $\sigma_i$ for every unit, and it is continuous in $\sigma_i$. For the thresholds (7) that are optimal for agreement between different ranking procedures, it also happens to land at position $H_\sigma \in \{x, \lambda\}$ separately for all possible values of $\sigma$. Induced by the marginal distribution of $(\theta, \lambda, \sigma)$, the critical distance among units,

$$V_\alpha(D_i) = P(\theta_i \geq \theta_\alpha | D_i)$$

$$\lambda_\alpha = H^{-1}_\alpha(1 - \alpha).$$

We also retain the assumption of mutual independence in Theorem 3. The critical distance is $\sigma_i$ for every unit, and it is continuous in $\sigma_i$. For the thresholds (7) that are optimal for agreement between different ranking procedures, it also happens to land at position $H_\sigma \in \{x, \lambda\}$ separately for all possible values of $\sigma$. Induced by the marginal distribution of $(\theta, \lambda, \sigma)$, the critical distance among units,

$$V_\alpha(D_i) = P(\theta_i \geq \theta_\alpha | D_i)$$

$$\lambda_\alpha = H^{-1}_\alpha(1 - \alpha).$$

We also retain the assumption of mutual independence in Theorem 3. The critical distance is $\sigma_i$ for every unit, and it is continuous in $\sigma_i$. For the thresholds (7) that are optimal for agreement between different ranking procedures, it also happens to land at position $H_\sigma \in \{x, \lambda\}$ separately for all possible values of $\sigma$. Induced by the marginal distribution of $(\theta, \lambda, \sigma)$, the critical distance among units,

$$V_\alpha(D_i) = P(\theta_i \geq \theta_\alpha | D_i)$$

$$\lambda_\alpha = H^{-1}_\alpha(1 - \alpha).$$

We also retain the assumption of mutual independence in Theorem 3. The critical distance is $\sigma_i$ for every unit, and it is continuous in $\sigma_i$. For the thresholds (7) that are optimal for agreement between different ranking procedures, it also happens to land at position $H_\sigma \in \{x, \lambda\}$ separately for all possible values of $\sigma$. Induced by the marginal distribution of $(\theta, \lambda, \sigma)$, the critical distance among units,

$$V_\alpha(D_i) = P(\theta_i \geq \theta_\alpha | D_i)$$

$$\lambda_\alpha = H^{-1}_\alpha(1 - \alpha).$$

We also retain the assumption of mutual independence in Theorem 3. The critical distance is $\sigma_i$ for every unit, and it is continuous in $\sigma_i$. For the thresholds (7) that are optimal for agreement between different ranking procedures, it also happens to land at position $H_\sigma \in \{x, \lambda\}$ separately for all possible values of $\sigma$. Induced by the marginal distribution of $(\theta, \lambda, \sigma)$, the critical distance among units,

$$V_\alpha(D_i) = P(\theta_i \geq \theta_\alpha | D_i)$$

$$\lambda_\alpha = H^{-1}_\alpha(1 - \alpha).$$

We also retain the assumption of mutual independence in Theorem 3. The critical distance is $\sigma_i$ for every unit, and it is continuous in $\sigma_i$. For the thresholds (7) that are optimal for agreement between different ranking procedures, it also happens to land at position $H_\sigma \in \{x, \lambda\}$ separately for all possible values of $\sigma$. Induced by the marginal distribution of $(\theta, \lambda, \sigma)$, the critical distance among units,
how to calculate r-values
Binomial likelihood

e.g., Beta prior/posteriors

Free Throw Ability

density
Binomial likelihood
e.g., Beta prior/posteriors
e.g., NBA

\[ P(\theta_i \geq \theta_\alpha | D_i) \]

two examples

\[ \lambda^\alpha_r \]

\[ D_{Ray.Allen} = 105/116 \]

\[ D_{LeBron.James} = 439/585 \]
<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>
Enhanced ranking and selection

a. MLE
b. p-value
c. posterior mean
d. maximal agreement

Fig. 3. Threshold functions, T2D example, data and axes as in Fig 1: Calculations use an inverse-gamma model for $\sigma^2$. For twenty threshold functions shown, ranging in $\alpha$ values from a small positive value (red) just including the first data point up to $\alpha = 0$ (blue). (Most data points are truncated by the plot, as in Fig 1; also, the grid is uniform on the scale of $\log \alpha$. 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.)
$\sigma^2 \sim \text{Exp}(1)$
RNAi example

MLE

p-value

posterior mean

$(X-R)/(X+R)$
Predictive accuracy

\[ E[ \text{similarity}_{t} \mid \text{Ranks}(\theta), \text{Ranks.hat[midseason]} ] \mid \text{complete season} ] \]

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

\[ t = \text{rank from top} \]
Bayes?

multi-loss

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

$$\text{risk}_\alpha = 1 - P \{ \delta(D_i) \leq \alpha, \theta \geq \theta_\alpha \},$$

marginally constrained

$$\text{risk}_\alpha + \gamma_\alpha P \{ \delta(D_i) \leq \alpha \}$$
Nested subsets of sample space: \( \{ \Gamma_\alpha : \alpha \in (0, 1) \} \)

<table>
<thead>
<tr>
<th>Constraint</th>
<th>( \inf { \alpha : D \in \Gamma_\alpha } )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( P(D \in \Gamma_\alpha</td>
<td>H_0) = \alpha )</td>
</tr>
<tr>
<td>( P(H_0</td>
<td>D \in \Gamma_\alpha) = \alpha )</td>
</tr>
<tr>
<td>( P(D \in \Gamma_\alpha) = \alpha )</td>
<td>r-value</td>
</tr>
</tbody>
</table>
does smoothing matter?

Normal/Normal

Gamma(.5,.5)

n=500
sampling performance does n matter?

Normal/Normal Gamma(.5,.5)

(a) $n=10000$

(b) $n=1000$

(c) $n=200$

(d) $n=50$
Theorem 4: If A1-A4, then as $n \to \infty$, any $\alpha$,

a. \[ \frac{1}{n} \sum_{i=1}^{n} 1 \left[ \hat{r}_n(X_i, \sigma_i^2) \leq \alpha \right] \longrightarrow_P \alpha \]

b. \[ \frac{1}{n} \sum_{i=1}^{n} 1 \left[ \hat{r}_n(X_i, \sigma_i^2) \leq \alpha \right] 1[\theta_i \geq \theta_{\alpha}] \longrightarrow_P A^*_\alpha \]
Sufficient conditions

A1. i.i.d. \((\theta_i, X_i, \sigma_i^2)\), \(\theta_i \sim f = F' \perp \sigma_i^2 \sim g\) \(f, g > 0\)

A2. \(\hat{F}_n \Rightarrow F\), a.s. \(\hat{F}_n\) permutation invariant

A3. \(p(x|\theta, \sigma^2)\) known form
   - continuous in \((x, \theta, \sigma^2)\)
   - positive, bounded in \((x, \theta)\) for each \(\sigma^2\)
   - monotone: for any \(x_1 > x_0\),
     \[\frac{p(x_1|\theta, \sigma^2)}{p(x_0|\theta, \sigma^2)}\] non-decreasing in \(\theta\)

A4. No crossing, limit case
Assume Normal/Normal

But really Normal/t

n=2000

what if model is wrong?

Sampling performance
sampling performance

what if model is wrong?

Assume Normal/Normal

But really Normal/mix

n=2000 (π₀, τ²)

\(3/4, 1/2\)

\(4/5, 1\)

\(3/4, 2\)

\(1/2, 1\)
what if model is wrong?

Assume Normal/Normal

But dependence

n=2000

ρ=0.19

ρ=0.24

ρ=0.55

ρ=0.71
• empirical Bayes ranking to maximize agreement

• setting: large-scale, non-sparse, differential variation

• less shrinkage than PM

• R-package at CRAN: rvalues
  Github: wiscstatman/rvalues