Controlling false discoveries in high dimensional situations:

Boosting with stability selection

Benjamin Hofner

Institut für Medizininformatik, Biometrie und Epidemiologie
Friedrich-Alexander-Universität Erlangen-Nürnberg

60. Biometrisches Kolloquium
Bremen - 2014

FRIEDRICH-ALEXANDER
UNIVERSITÄT
ERLANGEN-NÜRNBERG
MEDIZINISCHE FAKULTÄT

Model Fitting with Component-Wise Boosting

Linear Model

$$\mathbb{E}(y|\boldsymbol{x}_i) = \eta_i(\boldsymbol{x}_i)$$

with **linear** predictor

$$\eta_i(\boldsymbol{x}_i) = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij},$$

- Generally, model fitting aims at **minimizing the expected loss** with an appropriate **loss function** ρ.
- Here, the loss function is the **squared error loss**:

$$\rho(y, \boldsymbol{x}) = (y - \eta(\boldsymbol{x}))^2$$

- In other models, the loss could, e.g., be the negative log likelihood.
- In practice: Minimization of the **empirical risk**

$$n^{-1} \sum_{i=1}^{n} \rho(y_i, \eta_i(\boldsymbol{x}_i)) = n^{-1} \sum_{i=1}^{n} (y_i - \eta_i(\boldsymbol{x}_i))^2$$

Boosting

- minimizes empirical risk
- in a stagewise fashion
- via functional gradient descent (FGD).

In each iteration m

- the negative gradient of the loss function

$$u_i^{[m]} = - \left. \frac{\partial \rho(y_i, \eta)}{\partial \eta} \right|_{\eta = \hat{\eta}_i^{[m-1]}}$$

  is estimated via least squares base-learners ($\hat{\boldsymbol{u}}^{[m]} = \hat{\beta}_j \boldsymbol{x}_j$)
- only the model term corresponding to the **best-fitting base-learner** $\hat{\beta}_{j*}$ is updated by adding a **small fraction** ν **of the estimate** $\hat{\beta}_{j*}$ (e.g., 10%) to the model
- with "early stopping" variable selection is achieved

Practical notes

- We get an interpretable model, similar to models from maximum likelihood estimation or least squares estimation.
- Additionally, regularization is achieved via base-learner selection and shrinkage.

Practical notes

- We get an interpretable model, similar to models from maximum likelihood estimation or least squares estimation.
- Additionally, regularization is achieved via base-learner selection and shrinkage.

Yet,

- in high-dimensional settings, i.e., with many predictors, we might select a lot of uninformative variables.
- In many situations a **formal selection procedure with error control** seems advisable.

Stability Selection [Meinshausen and Bühlmann 2010]

... is a versatile approach, which can be combined with (all) high-dimensional variable selection approaches.

... is based on subsampling (▶ draw samples without replacement).

... controls the per-family error rate PFER $= \mathbb{E}(V)$, where V is the number of false positives.

	Keep H_0	Reject H_0	
H_0 true	U	V	m_0
H_1 true	T	S	$m - m_0$
	$m - R$	R	m

Overview: Error Rates [see e.g. Dudoit et al. 2003]

per-comparison error rate (PCER)[1]: $\mathbb{E}(V)/m$

standard testing procedure, no multiplicity correction

per-family error rate (PFER): $\mathbb{E}(V)$

family-wise error rate (FWER): $\mathbb{P}(V \geq 1)$

false discovery rate (FDR)[1]*:* $\mathbb{E}(\frac{V}{R})$

	Keep H_0	Reject H_0	
H_0 true	U	V	m_0
H_1 true	T	S	$m - m_0$
	$m - R$	R	m

[1] m = number of tested hypothesis, R = number of rejected hypothesis

Overview: Error Rates [see e.g. Dudoit et al. 2003]

per-comparison error rate (PCER)[1]: $\mathbb{E}(V)/m$

standard testing procedure, no multiplicity correction

per-family error rate (PFER): $\mathbb{E}(V)$

family-wise error rate (FWER): $\mathbb{P}(V \geq 1)$

false discovery rate (FDR)[1]*:* $\mathbb{E}(\frac{V}{R})$

$$\text{PCER} \leq \text{FWER} \leq \text{PFER}$$

- For fixed α, PFER is more conservative than FWER
  FWER is more conservative than PCER

$$\text{FDR} \leq \text{FWER}$$

- For fixed α, FWER is more conservative than FDR

[1] m = number of tested hypothesis, R = number of rejected hypothesis

Stability Selection

Algorithm (simplified)

1. Select a random subset of size $\lfloor n/2 \rfloor$ of the data.
2. Fit boosting model until q variables are selected (out of p).
3. Record which variables were selected.
4. Repeat $B = 100$ times.

5. Compute selection frequency per variable.
6. Select variables with frequency $\geq \pi_{\text{thr}}$.

Stability Selection

Algorithm (simplified)

1. Select a random subset of size $\lfloor n/2 \rfloor$ of the data.
2. Fit boosting model until q variables are selected (out of p).
3. Record which variables were selected.
4. Repeat $B = 100$ times.

5. Compute selection frequency per variable.
6. Select variables with frequency $\geq \pi_{\text{thr}}$.

- Upper bound for the per-family error rate (PFER) is given by:

$$\mathbb{E}(V) \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p}$$

(if the exchangeability assumption holds for all noise variables)

Improved Stability Selection [Shah and Samworth 2013]

- Use complementary pairs (use subsample of size $\lfloor n/2 \rfloor$ and its complement) ▶ $2B$ subsamples
- Tighter, i.e., less conservative error bounds can be derived under certain conditions. This leads to a less conservative selection procedure:

  a) **If distribution of simultaneous selection probabilities is unimodal:**

$$\mathbb{E}(V) \leq \frac{q^2}{c(\pi_{\text{thr}}, B) \cdot p} \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p}$$

  b) **If distribution of (simultaneous) selection probabilities is r-concave:**

$$\begin{aligned} \mathbb{E}(V) &\leq \min\left\{ D\left(2\pi_{\text{thr}} - 1; \frac{q^2}{p^2}, B, -\frac{1}{2}\right), D\left(\pi_{\text{thr}}; \frac{q}{p}, 2B, -\frac{1}{4}\right)\right\} p \\ &\leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p} \end{aligned}$$

  with functions $c(\cdot, \cdot)$ and $D(\cdot, \cdot, \cdot, \cdot)$.

Improved Stability Selection [Shah and Samworth 2013]

- Use complementary pairs (use subsample of size $\lfloor n/2 \rfloor$ and its complement) ▶ $2B$ subsamples
- Tighter, i.e., less conservative error bounds can be derived under certain conditions. This leads to a less conservative selection procedure:

  a) **If distribution of simultaneous selection probabilities is unimodal:**

$$\mathbb{E}(V) \leq \frac{q^2}{c(\pi_{\text{thr}}, B) \cdot p} \leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p}$$

  b) **If distribution of (simultaneous) selection probabilities is r-concave:**

$$\begin{aligned} \mathbb{E}(V) &\leq \min\left\{ D\left(2\pi_{\text{thr}} - 1; \frac{q^2}{p^2}, B, -\frac{1}{2}\right), D\left(\pi_{\text{thr}}; \frac{q}{p}, 2B, -\frac{1}{4}\right)\right\} p \\ &\leq \frac{q^2}{(2\pi_{\text{thr}} - 1)p} \end{aligned}$$

  with functions $c(\cdot, \cdot)$ and $D(\cdot, \cdot, \cdot, \cdot)$.

Condition b) is stronger than a) and might not always hold, especially for larger numbers of subsamples B. Thus a) is usually recommended.

Simulation Study

Question: How does the data set and the choice of tuning parameters influence the performance of stability selection?

Settings:

- Sample size n: 50, 100, 500
- Number of variables p: 100, 500, 1000
- Number of influential variables p_{infl}: 2, 3, 8

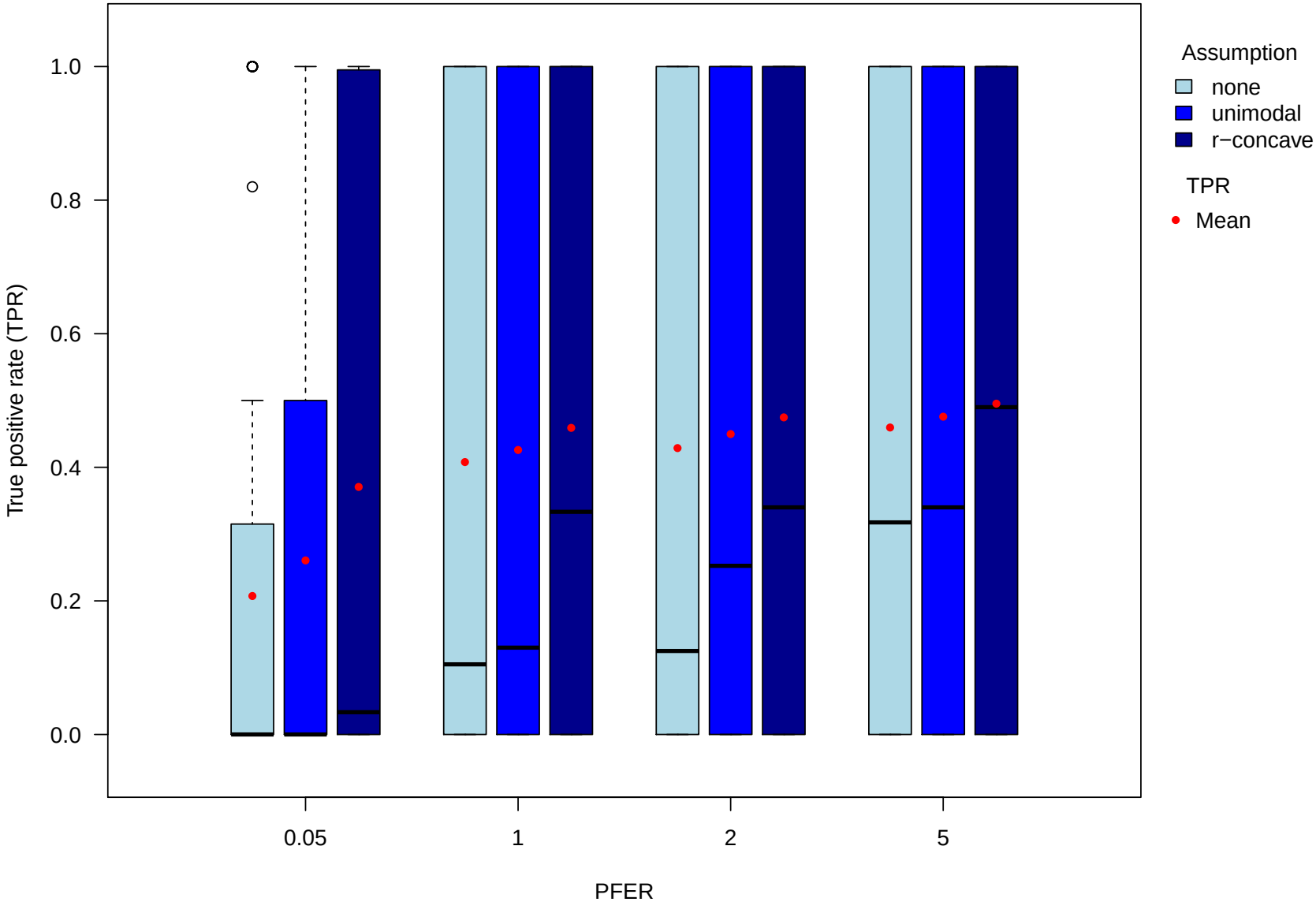
- Threshold π_{thr}: 0.6, 0.75, 0.9
- Upper bound for PFER: 0.05, 1, 2, 5

- All three error bounds were used.
- $B = 100$ ($B = 50$ for complementary pairs subsampling)

Each setting was repeated 50 times.

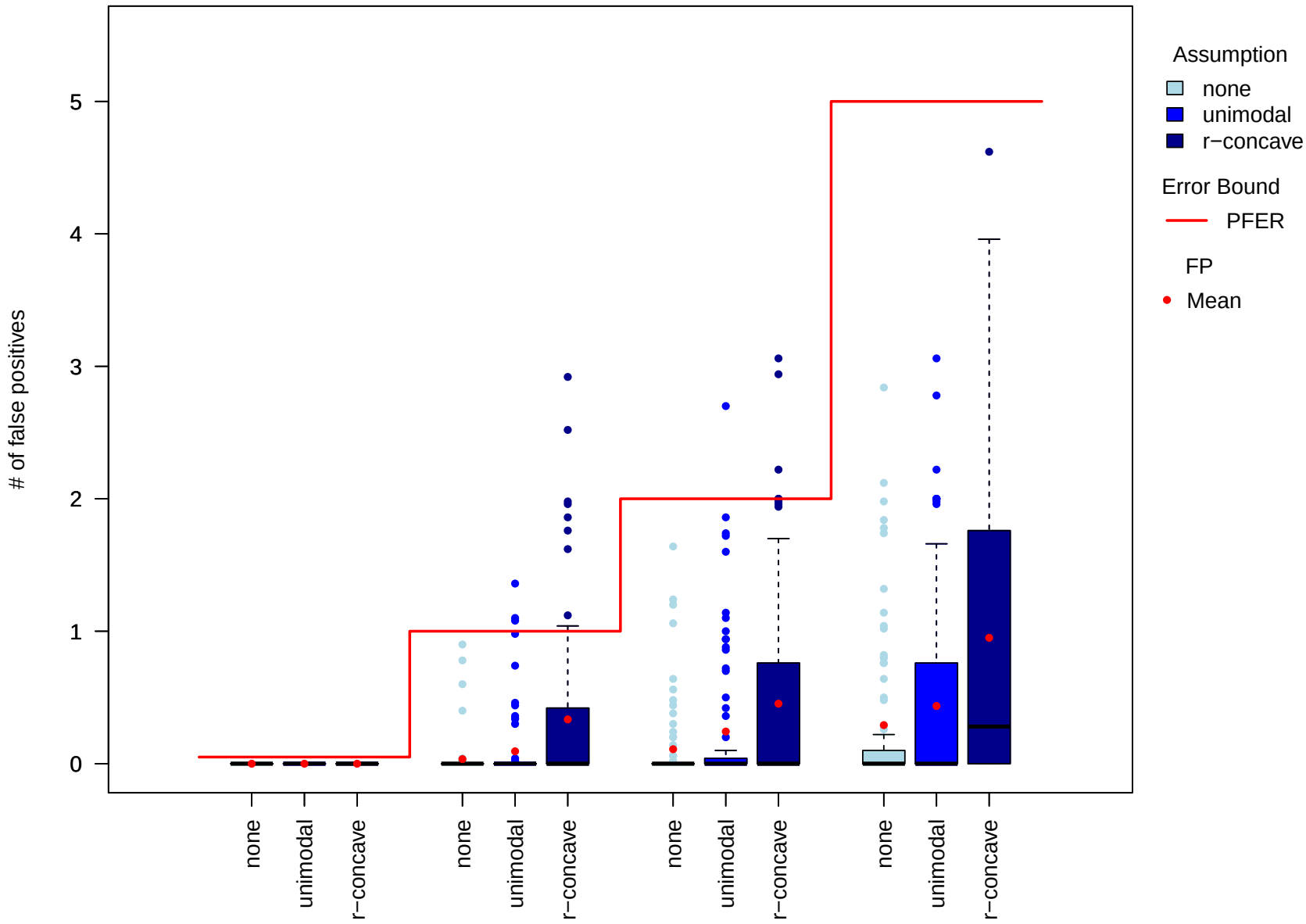
Simulation Study

Influence of Assumptions on the TP rate



Simulation Study

Influence of Assumptions on the number of FP



Summary of Simulation Results

- PFER is conservatively controlled
  (less conservative if tighter error bounds are used)
- If tighter error bounds are use: some violations of error bounds
  - no assumption: 0 % (0)
  - unimodal: 1.2 % (4)
  - r-concave: 4.0 % (13)

  (assumptions seem to be violated)
- In general, stability selection is very conservative
  (as it controls the PFER)

- TPR decreases with increasing number of influential variables p_{infl}
- TPR increases with increasing number of observations n

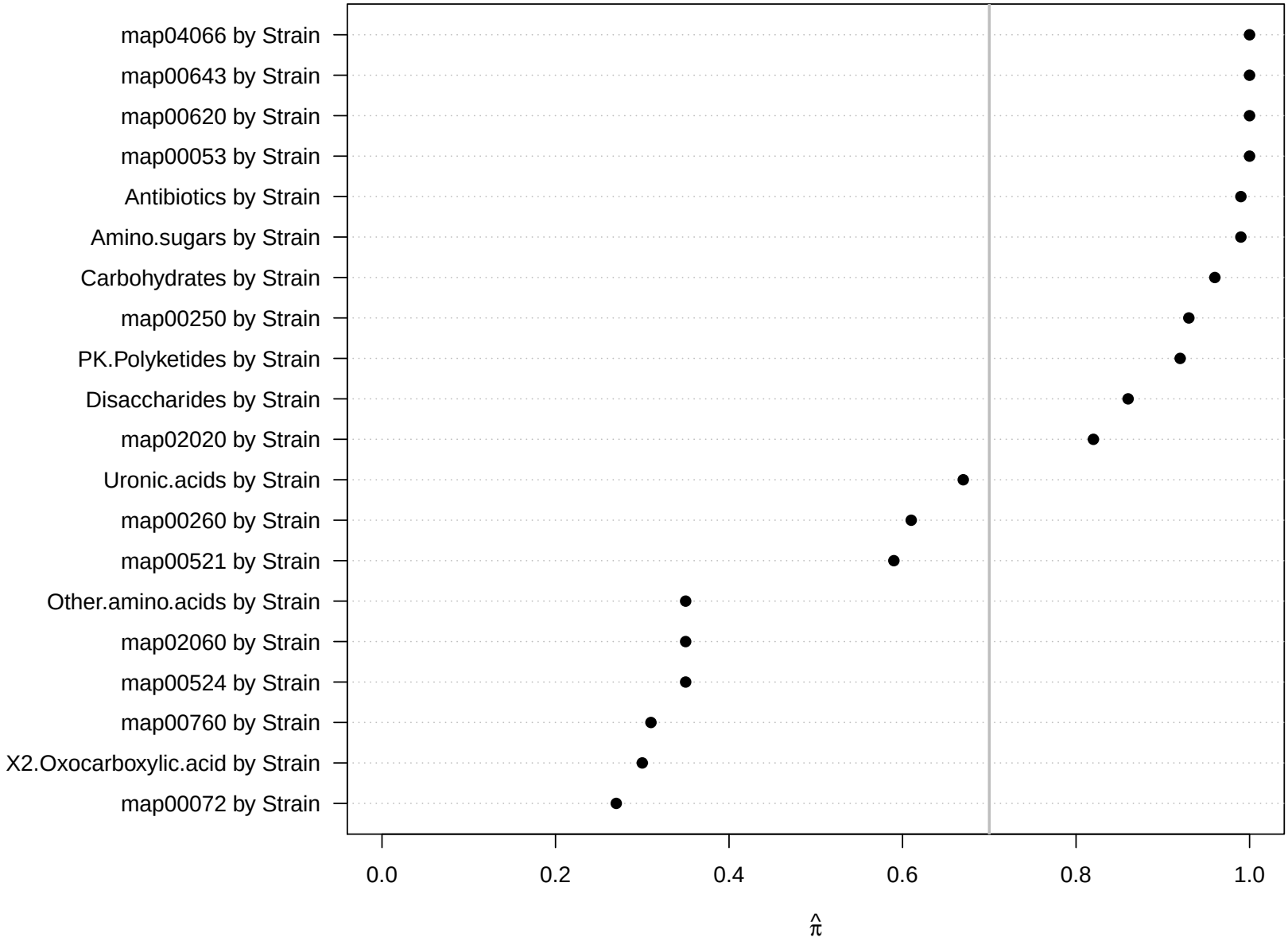
- Choice of threshold π_{thr} is less important
  (as long as it is large enough to result in enough variables q to be selected)

E. coli data

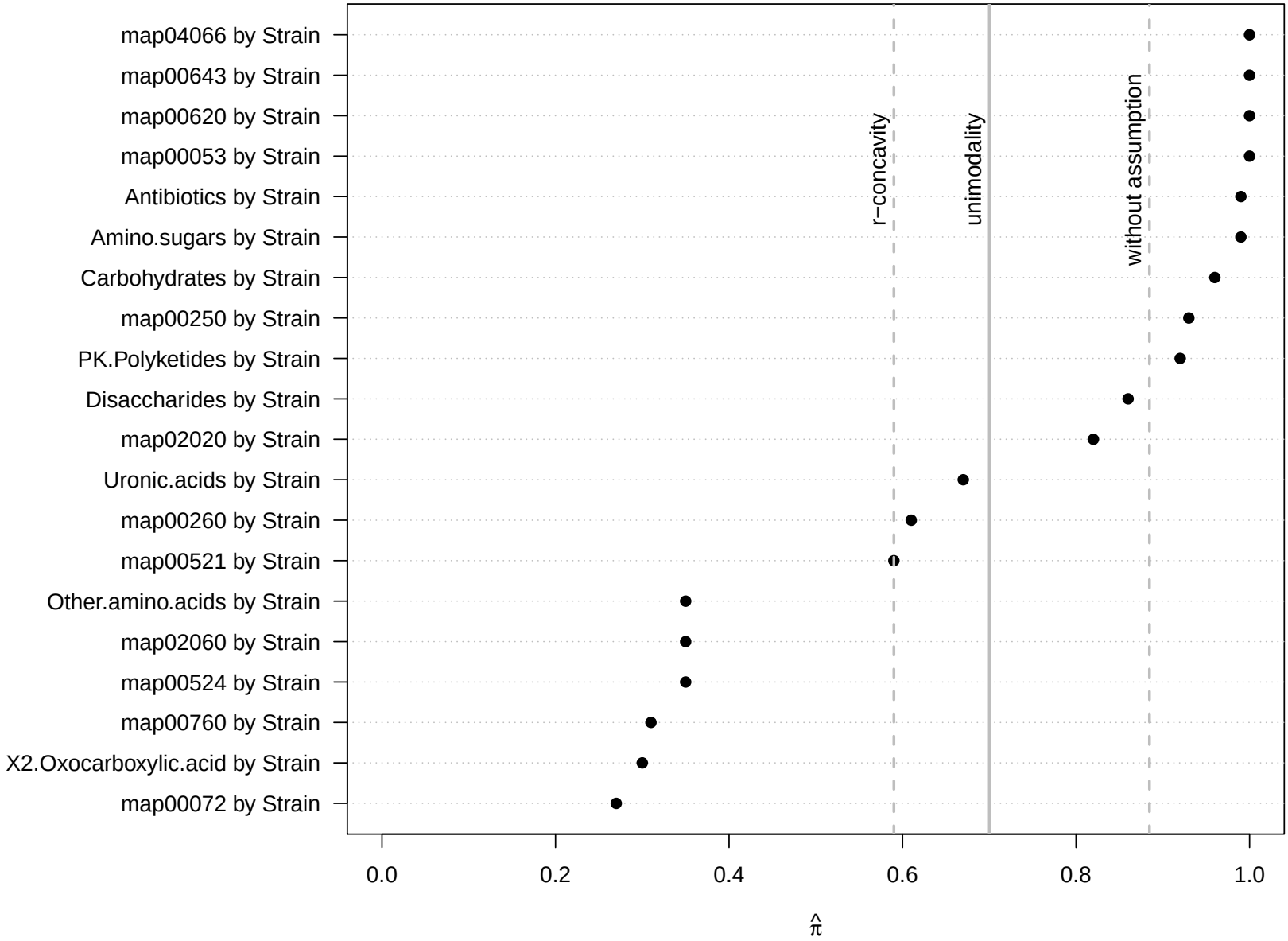
- Data set based on `vaas_et_al` from package **opmdata** [Vaas et al. 2012; 2013]
- **Aim:** Find metabolic pathways / substrate groups that differ between two *E. coli* strains
- Use 97 pathway annotations / substrate groups as main effects and strain specific effects as possible predictors (▶ 195 predictors in total)
- Use offset model with main effects only and check for additional predictors.

- We use an upper bound for the PFER $= 1.5$, i.e. we accept **at maximum** 1.5 (expected) false discoveries.
- Additionally, the number of included variables per boosting model was set to $q = 15$ (which results in a cutoff $\pi_{\text{thr}} = 0.7$).
- We used the unimodality assumption to compute the error bound.

Results: E. coli data



Results: E. coli data



Results: E. coli data

- Number of selected base-learners with standard cross-validation: 47.

- Stability selection:
  - 11 strain specific pathways / substrate groups were identified as stable effects.
  - Without further assumptions, 9 effects were considered stable.
  - With r-concavity assumption, 14 effects where considered stable.

  - In all settings, the number of influential variables is way above the PFER of 1.5 and much lower than the cross-validation result.

Summary and Outlook

- Stability selection works well in conjunction with boosting.
- It is especially useful in sparse, high-dimensional settings
- and it controls the PFER.
- Stability selection results in a fundamentally new solution which cannot be recreated elsewise (e.g. by selecting a certain regularization parameter, here m_{stop}).
- Stability selection can also be used for boosted GAM models and other fitting approaches.

- Yet, stability selection is quite conservative
  - as it controls the PFER
  - and as even this control seems to be conservative (at least for the standard error bound).
- Higher selection numbers (i.e. higher TPR) can be obtained by tighter error bounds, yet, sometimes error bounds do not hold any more.

Summary and Outlook

- Stability selection is implemented in the *R* package **mboost** [Hothorn et al. 2013, Hofner et al. 2014] in the function `stabsel()`.

- Tighter error bounds are currently only implemented in the development version **mboostDevel** on `http://r-forge.r-project.org/projects/mboost/`.
- The latter package also implements
  `stabsel_parameters(cutoff, q, PFER)`
  a function to compute error bounds for combinations of two of the three parameters.
- Stability selection can also be used with other fitting functions.

- Eventually, these new/improved functions will be migrated to **mboost**.

References

Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18:71–103, 2003.

Benjamin Hofner, Andreas Mayr, Nicolay Robinzonov, and Matthias Schmid. Model-based boosting in R – A hands-on tutorial using the R package mboost. *Computational Statistics*, 29:3–35, 2014.

Torsten Hothorn, Peter Bühlmann, Thomas Kneib, Matthias Schmid, and Benjamin Hofner. *mboost: Model-Based Boosting*, 2013. URL `http://CRAN.R-project.org/package=mboost`. R package version 2.2-3.

Nicolai Meinshausen and Peter Bühlmann. Stability selection (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72: 417–473, 2010.

Rajen D. Shah and Richard J. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75:55–80, 2013.

Lea A. I. Vaas, Johannes Sikorski, Victoria Michael, Markus Göker, and Hans-Peter Klenk. Visualization and curve-parameter estimation strategies for efficient exploration of phenotype microarray kinetics. *PLoS ONE*, 7(4):e34846, 2012.

Lea A. I. Vaas, Johannes Sikorski, Benjamin Hofner, Nora Buddruhs, Anne Fiebig, Hans-Peter Klenk, and Markus Göker. opm: An r package for analysing omnilog®phenotype microarray data. *Bioinformatics*, 29(14):1823–1824, 2013.