Sequential/Adaptive Benchmarking

Manuel J. A. Eugster

Institut für Statistik
Ludwig-Maximiliams-Universität München

Forschungskolloqium IMBS, Universität zu Lübeck, 2010

Hauptmenü

- Home
- Anreise
- Mitarbeiter
- Weblinks

Forschung

- Auszeichnungen
- Projekte
- Publikationen
- Software
- Stellenausschreibung

Kolloquium

Datum	Vortragender	Titel
26. November 2010	Prof. Dr. Lothar Kreienbrock	BIOMETRISCHE AUFGABEN IM BEREICH VETERINARY...
19. November 2010	Manuel J. A. Eugster (Dipl.-Ing.)	SEQUENTIAL/ADAPTIVE BENCHMARKING
29. Oktober 2010	Dr. Jochem König	VERALLGEMEINERTE LOGISTISCHE REGRESSION IN DER...
27. Oktober 2010	Dr. Nele Friedrich	SURVIVALANALYSEN: VERGLEICH VERSCHIEDENER PRÄ...
22. Oktober 2010	Bernd Bischl (M.Sc.)	BENCHMARKING AND ANALYSIS OF LOCAL CLASSIFICATION...
2. Juli 2010	Dr. Andreas Faldum	Adaptiv-sequentielle Studien mit adäquater Berü...
4. Juni 2010	Dr. Rebecca Hein	Adaption of an Empirical Bayes approach to...

Hauptmenü

- Home
- Anreise
- Mitarbeiter
- Weblinks

Forschung

- Auszeichnungen
- Projekte
- Publikationen
- Software
- Stellenausschreibung

Kolloquium

Datum	Vortragender	Titel
26. November 2010	Prof. Dr. Lothar Kreienbrock	BIOMETRISCHE AUFGABEN IM BEREICH VETERINARY...
19. November 2010	Manuel J. A. Eugster (Dipl.-Ing.)	SEQUENTIAL/ADAPTIVE BENCHMARKING
29. Oktober 2010	Dr. Jochem König	VERALLGEMEINERTE LOGISTISCHE REGRESSION IN DER...
27. Oktober 2010	Dr. Nele Friedrich	SURVIVALANALYSEN: VERGLEICH VERSCHIEDENER PRÄ...
22. Oktober 2010	Bernd Bischl (M.Sc.)	BENCHMARKING AND ANALYSIS OF LOCAL CLASSIFICATION
2. Juli 2010	Dr. Andreas Faldum	Adaptiv-sequentielle Studien mit adäquater Berü...
4. Juni 2010	Dr. Rebecca Hein	Adaption of an Empirical Bayes approach to...

Benchmark experiments

"Statistical decathlon": The modern decathlon (*benchmark experiment*) is a set combination of athletic disciplines (*data sets*), testing an athlete's (*learning algorithm's*) strength, speed, stamina, endurance and perseverance (*performance measures*).

(*) Based on Wikipedia (2010).

Data generating process:

Given a data generating process DGP, we draw B independent and identically distributed learning samples:

$$\mathfrak{L}^1 = \{z_1^1, \ldots, z_n^1\} \sim DGP$$

$$\vdots$$

$$\mathfrak{L}^B = \{z_1^B, \ldots, z_n^B\} \sim DGP$$

Candidate algorithms:

There are $K > 1$ algorithms a_k $(k = 1, \ldots, K)$ with the function $a_k(\cdot \mid \mathfrak{L}^b)$ the fitted model on the learning sample $\mathfrak{L}^b$.

(*) Following Hothorn, Leisch, Zeileis, and Hornik (2005).

Performance measure:

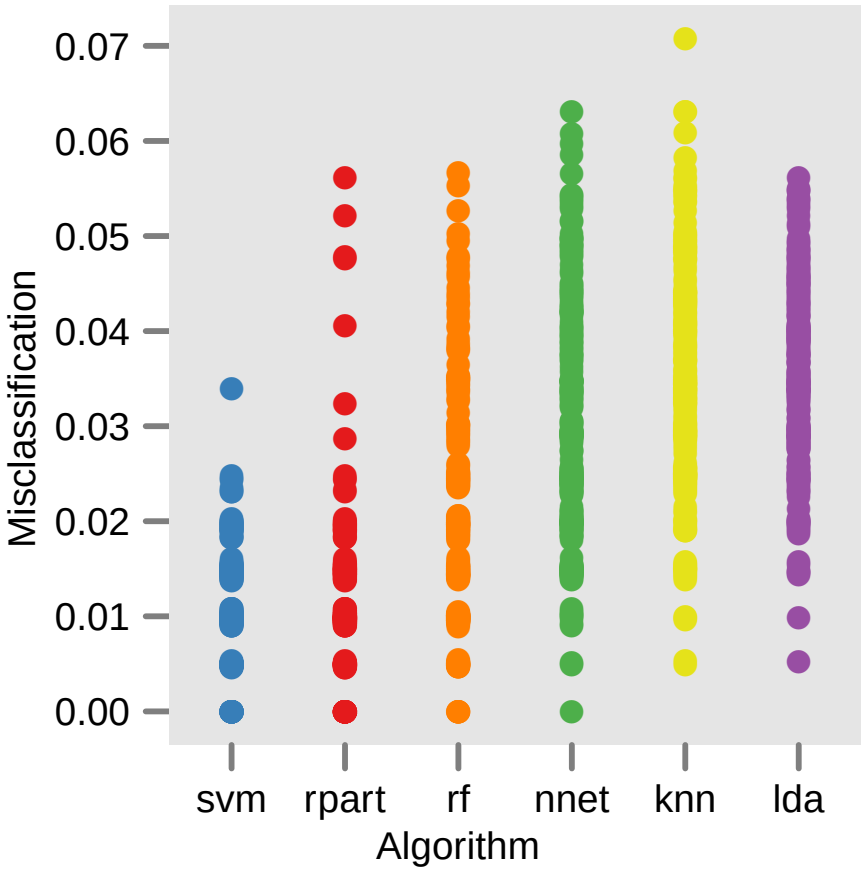
The performance of algorithm a_k when provided with the learning sample $\mathfrak{L}^b$ is measured by a scalar function p:

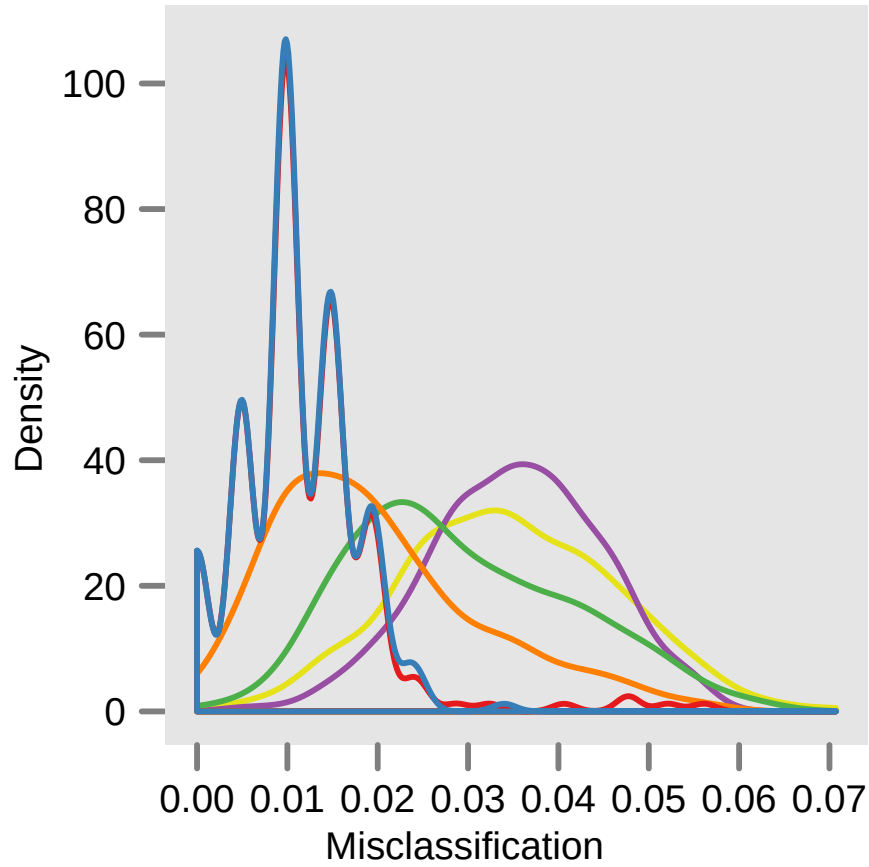
$$p_{kb} = p(a_k, \mathfrak{L}^b) \sim \mathcal{P}_k = \mathcal{P}_k(DGP)$$

Generalization performance:

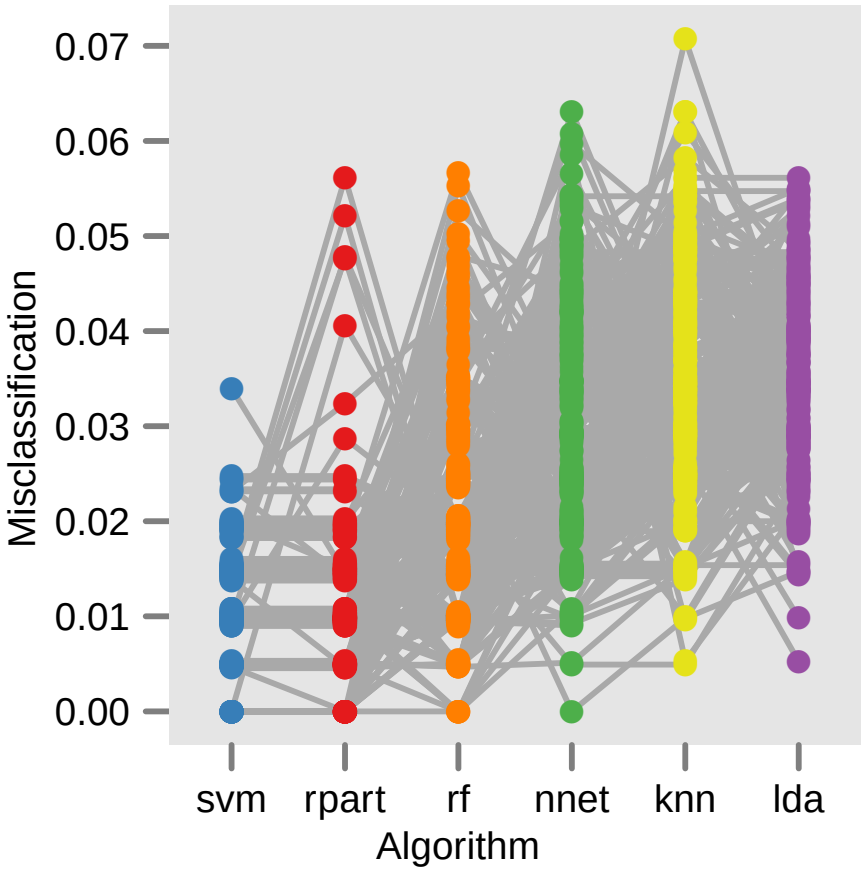
Empirical performance measure based on a validation sample $\mathfrak{T} \sim DGP$:

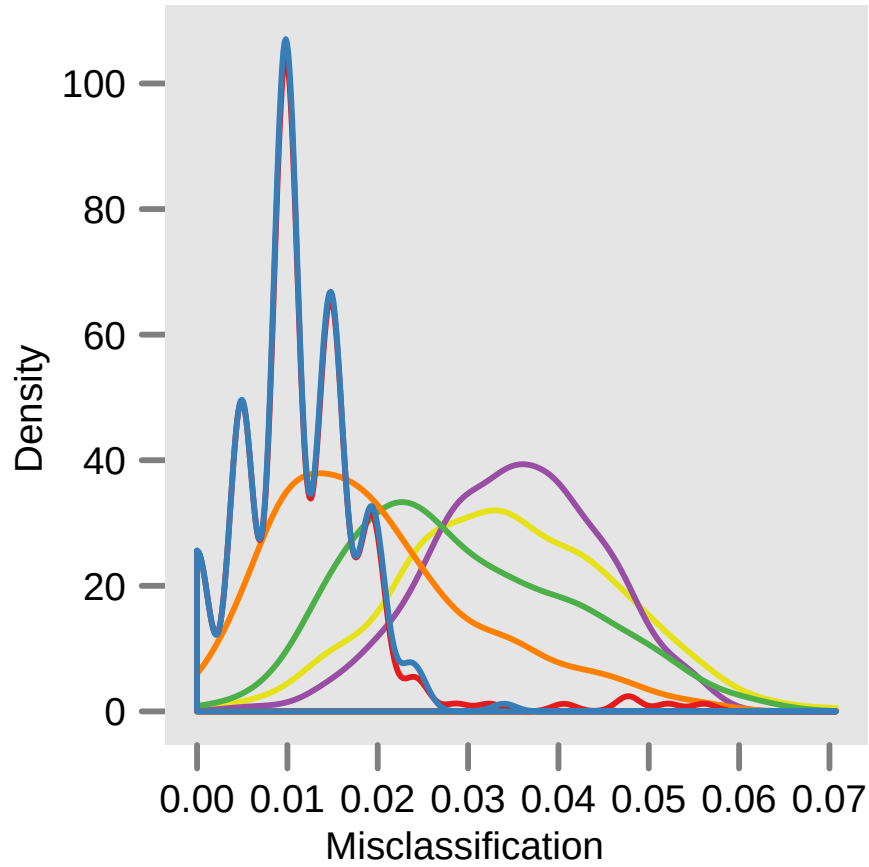
$$\hat{p}_{kb} = \hat{p}_{\mathfrak{T}}(a_k, \mathfrak{L}^b) \sim \hat{\mathcal{P}}_k = \hat{\mathcal{P}}_k(DGP)$$





Classification problem `monks3` with $B = 250$ replications, bootstrapping as resampling scheme to generate the learning samples $\mathfrak{L}^b$, and the out-of-bag scheme for $\mathfrak{T}^b$.





Classification problem `monks3` with $B = 250$ replications, bootstrapping as resampling scheme to generate the learning samples $\mathfrak{L}^b$, and the out-of-bag scheme for $\mathfrak{T}^b$.

Inference:

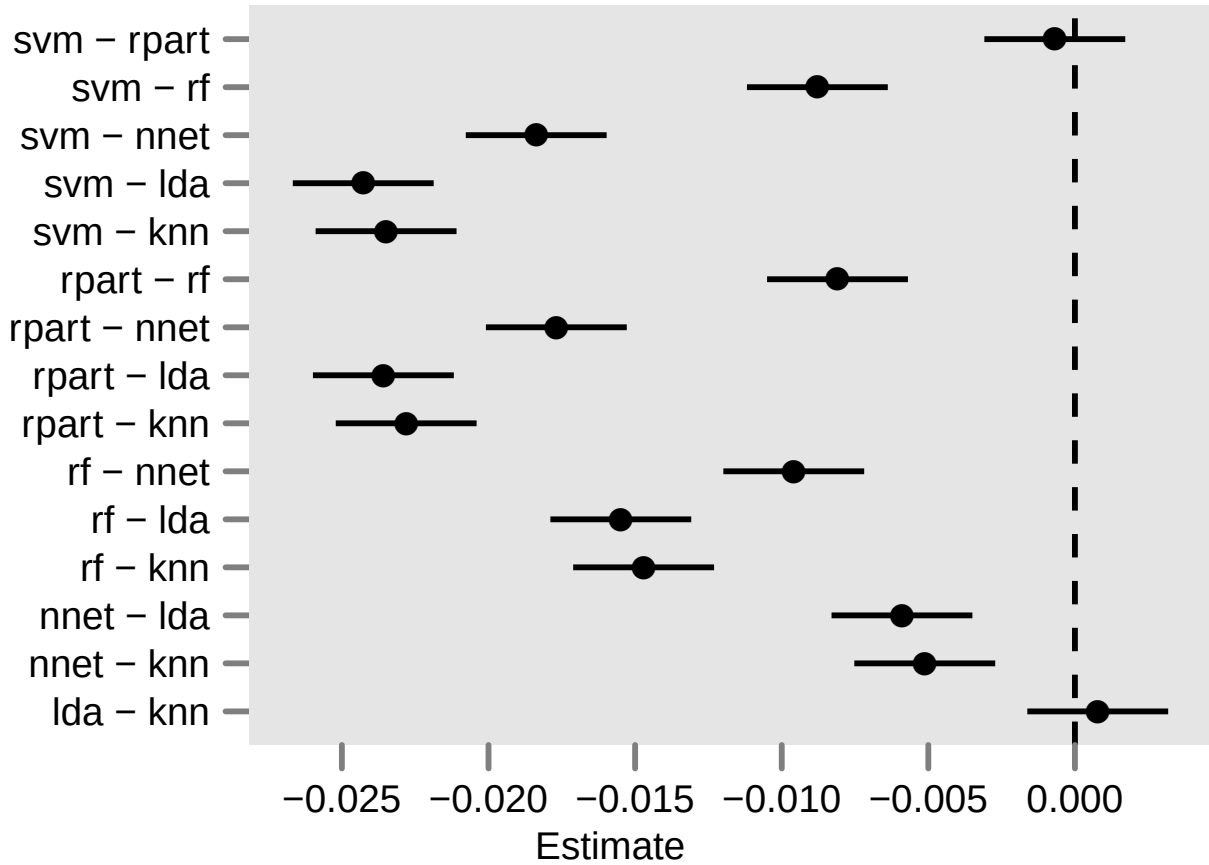
Given the K different random samples $\{\hat{p}_{k1}, \ldots, \hat{p}_{kB}\}$ with B iid samples drawn from the distributions $\hat{\mathcal{P}}_k(DGP)$ the null hypothesis of interest for most problems is:

$$H_0 \ : \ \hat{\mathcal{P}}_1 = \cdots = \hat{\mathcal{P}}_K$$

Test procedure:

An algorithm a_k is better than an algorithm $a_{k'}$ with respect to a performance measure p and a functional ϕ iff $\phi(\hat{\mathcal{P}}_k) < \phi(\hat{\mathcal{P}}_{k'})$ $(k, k' \in \{1, \ldots, K\})$.

$$T \begin{cases} H_0 : & \phi(\hat{\mathcal{P}}_1) = \cdots = \phi(\hat{\mathcal{P}}_K) \\ H_1 : & \exists k, k' : \phi(\hat{\mathcal{P}}_k) \neq \phi(\hat{\mathcal{P}}_{k'}) \end{cases}$$



Pairwise test decisions based on the 95% simultaneous confidence intervals computed for a linear mixed-effects model of the misclassification error using Tukey contrasts.

Preference relation:

$a_k \prec a_{k'}$ – algorithm a_k performs better than $a_{k'}$

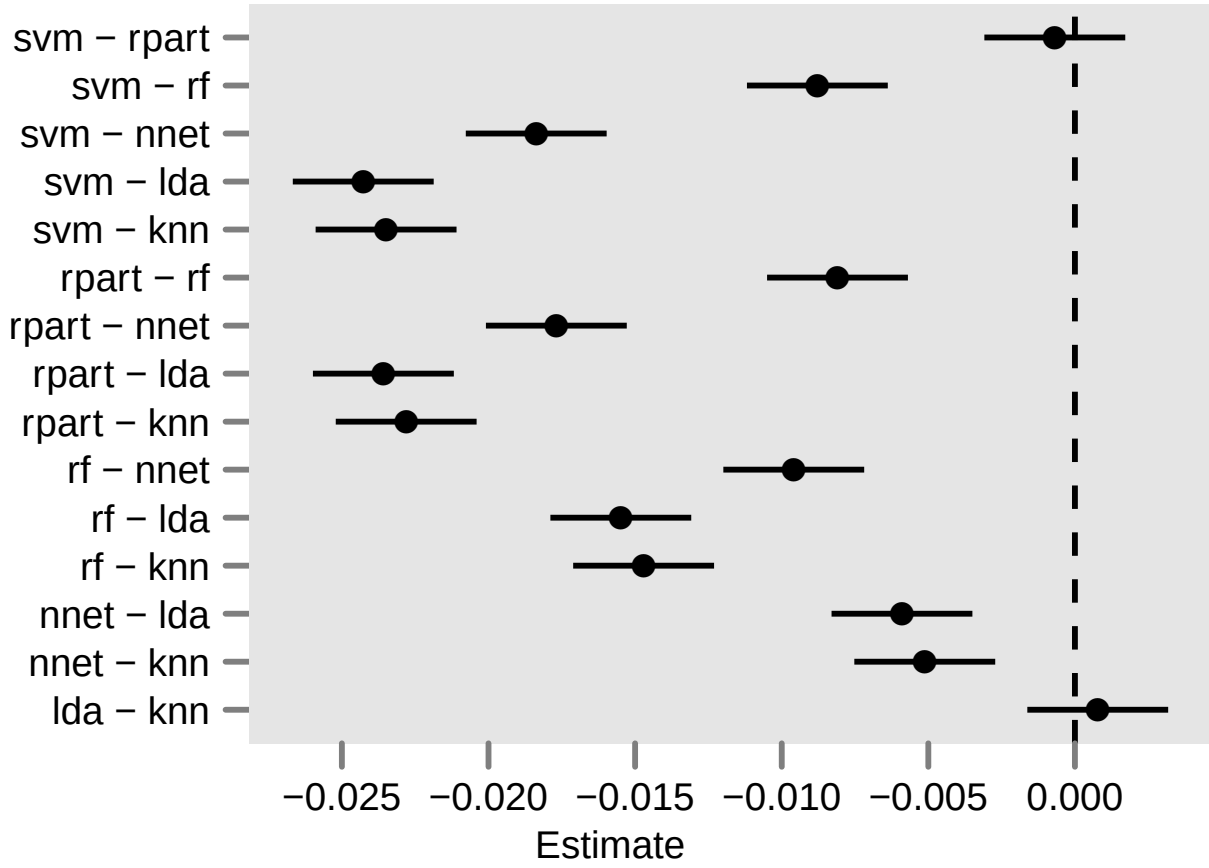
$a_k \sim a_{k'}$ – algorithm a_k performs equally to $a_{k'}$

An arbitrary pairwise comparison induces a mathematical relation R which we interpret as preference relation:

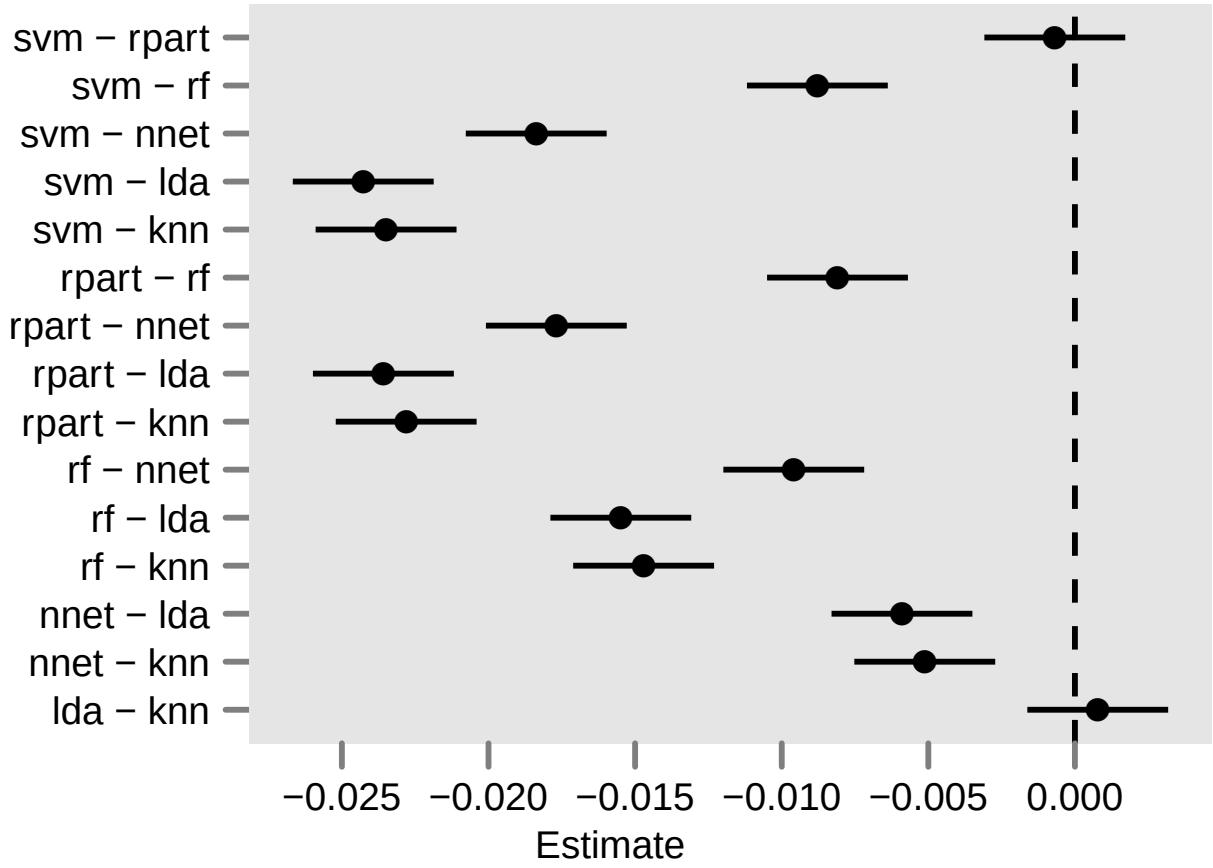
$$(a_k \; R \; a_{k'}) \Rightarrow a_k \sim a_{k'}$$

or

$$(a_k \; R \; a_{k'}) \Rightarrow a_k \prec a_{k'}$$



Relation R is "significantly different": (`svm` R `rf`), (`svm` R `nnet`), (`svm` R `lda`), ...



Relation R is "significantly different": (`svm` R `rf`), (`svm` R `nnet`), (`svm` R `lda`), ...

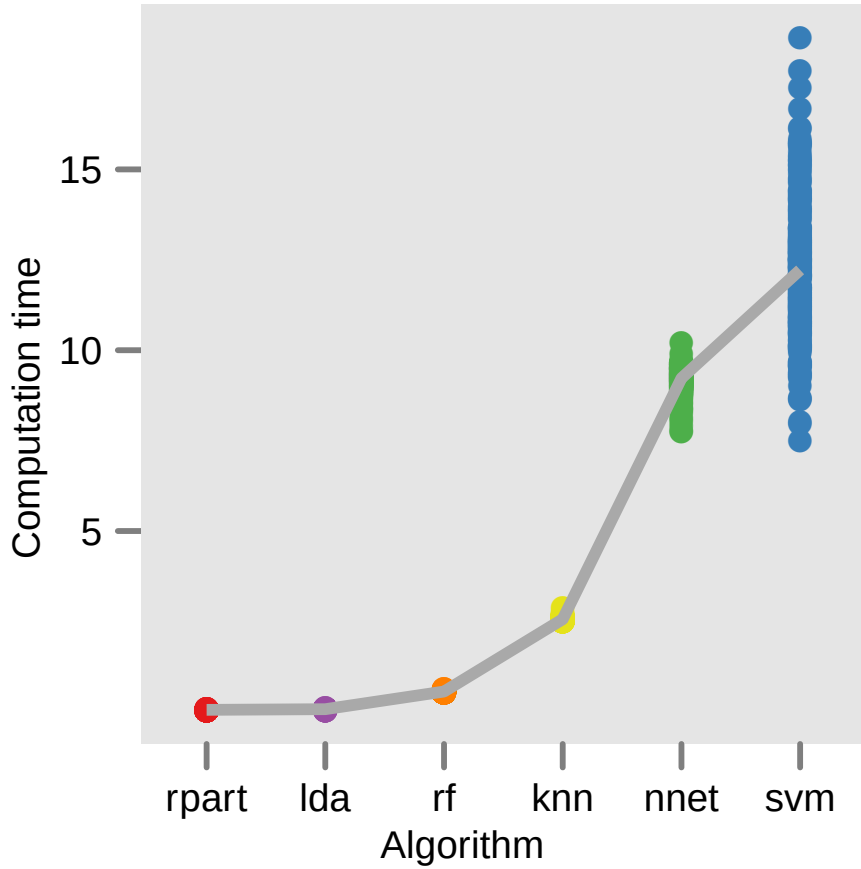
Preference relation (strict part):

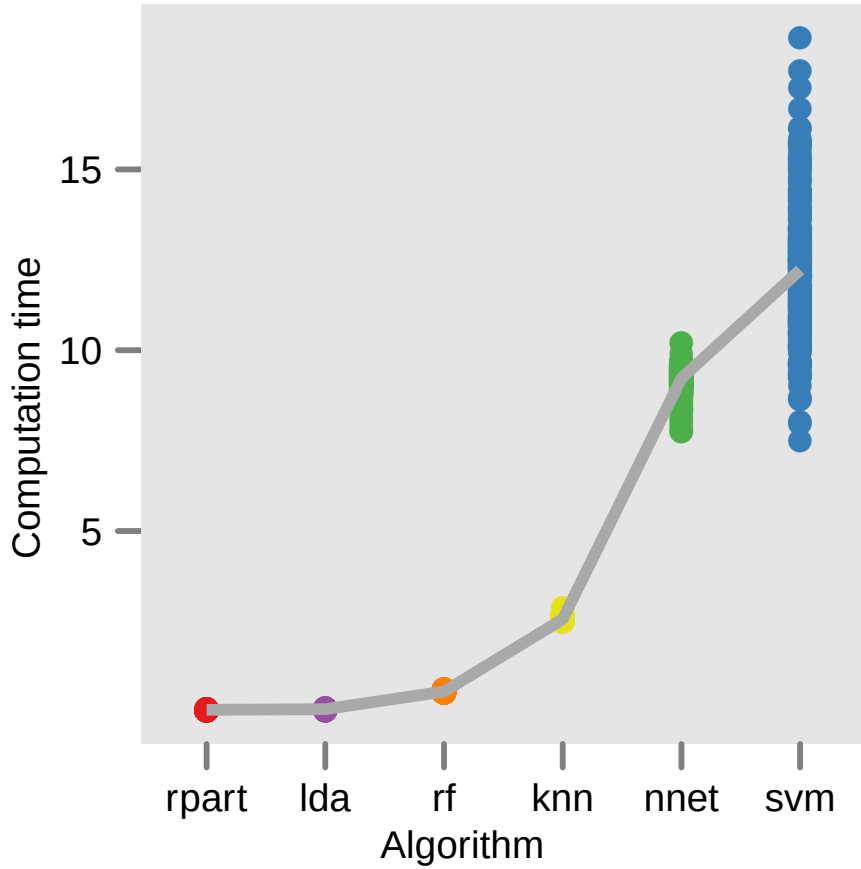
$$\texttt{svm} \sim \texttt{rpart} \prec \texttt{rf} \prec \texttt{nnet} \prec \texttt{knn} \sim \texttt{lda}$$

Preference combination:
Aggregate an ensemble of preference relations, each based on an performance measure of interest, using consensus decision-making methods:

$$\{R_1, \ldots, R_j\} \Rightarrow_w \bar{R}$$

Combination methods are, for example, Borda count, Condorcet approaches, optimization methods.


15
10
5
Computation time
rpart
lda
rf
knn
nnet
svm
Algorithm



Computation time ($w = 0.2$):	$\texttt{rpart} \prec \texttt{lda} \prec \texttt{rf} \prec \texttt{knn} \prec \texttt{nnet} \prec \texttt{svm}$
Misclassification ($w = 1$):	$\texttt{svm} \sim \texttt{rpart} \prec \texttt{rf} \prec \texttt{nnet} \prec \texttt{knn} \sim \texttt{lda}$
Consensus (linear order):	$\texttt{rpart} \prec \texttt{svm} \prec \texttt{rf} \prec \texttt{nnet} \prec \texttt{lda} \sim \texttt{knn}$

Sound benchmark experiment framework to compute a statistically correct order of the candidate algorithms, but ...

... it is based on a, in most benchmark experiments, **freely chosen** number of replications B:

$$\mathfrak{L}^1 = \{z_1^1, \ldots, z_n^1\} \sim DGP$$

$$\vdots$$

$$\mathfrak{L}^B = \{z_1^B, \ldots, z_n^B\} \sim DGP$$

(*) To be honest, in most cases B is specified based on the algorithms' running times.

For $b = 1, \ldots, B$

1. Draw learning sample $\mathfrak{L}^b$.
2. Measure performance p_{kb} of the $k = 1, \ldots, K$ candidate algorithms.

Execute test procedure T on the K performance estimations $\{p_{1k}, \ldots, p_{Bk}\}$ and make a decision for a given α.

For $b = 1, \ldots, B$

1. Draw learning sample $\mathfrak{L}^b$.
2. Measure performance p_{kb} of the $k = 1, \ldots, K$ candidate algorithms.

Execute test procedure T on the K performance estimations $\{p_{1k}, \ldots, p_{Bk}\}$ and make a decision for a given α.

- Benchmark experiments are considered as **fixed-sample** experiments; hypotheses of interests are tested using a test T at the end.
- The nature of benchmark experiments is **sequential**.

Do

1. Draw learning sample $\mathfrak{L}^b$.
2. Measure performance p_{bk} of the $k = 1, \ldots, K$ candidate algorithms.
3. Execute test procedure T on the K performance estimations $\{p_{1k}, \ldots, p_{bk}\}$.

While no decision for a given α (and $b \leq B$).

Do

1. Draw learning sample $\mathfrak{L}^b$.
2. Measure performance p_{bk} of the $k = 1, \ldots, K$ candidate algorithms.
3. Execute test procedure T on the K performance estimations $\{p_{1k}, \ldots, p_{bk}\}$.

While no decision for a given α (and $b \leq B$).

- **Sequential/Adaptive benchmarking:** execute test T successively on the accumulating data.
- This enables
  (1) to monitor the benchmark experiment, and
  (2) to make a decision – to *stop* or to *go on*.

Exemplar benchmark experiments

(1) $\mathfrak{L}$ is the Pima Indians Diabetes data set; **(2)** $\mathfrak{L}^b$ by bootstrapping; **(3)** linear discriminant analysis (`lda`), support vector machine with $C = 1.00$ (`svm1`), support vector machine with $C = 1.01$ (`svm2`), random forest (`rf`); **(4)** misclassification on the out-of-bag samples; **(5)** $B = 100$.

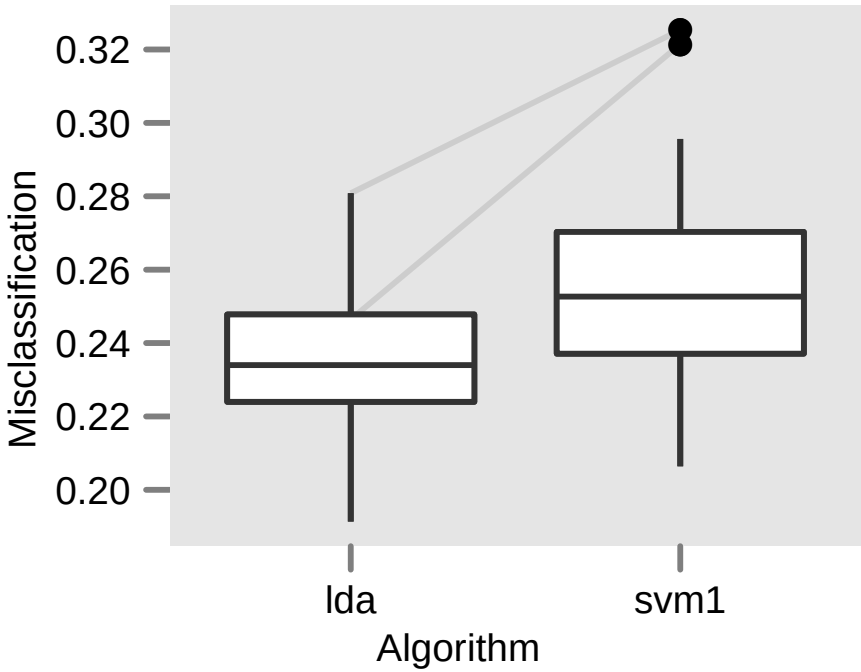
$\Rightarrow$ compare two algorithms at a time, i.e., test if algorithm a_1 is better than algorithm a_2.

(6) Wilcoxon Signed Rank test, $\alpha = 0.05$.

Monitoring

Observe and interpret the test result, mainly the p-value, on the accumulating performance measures.

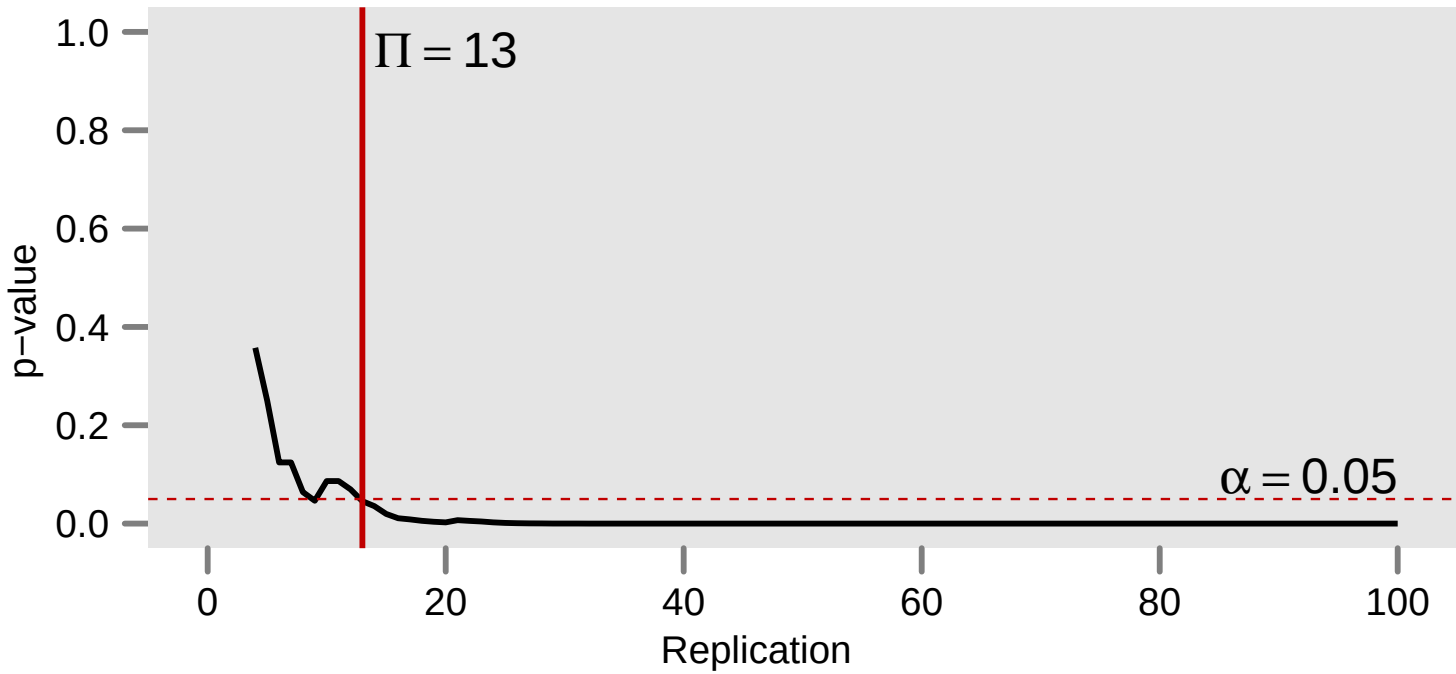
Scenario 1 – Different algorithm performances:



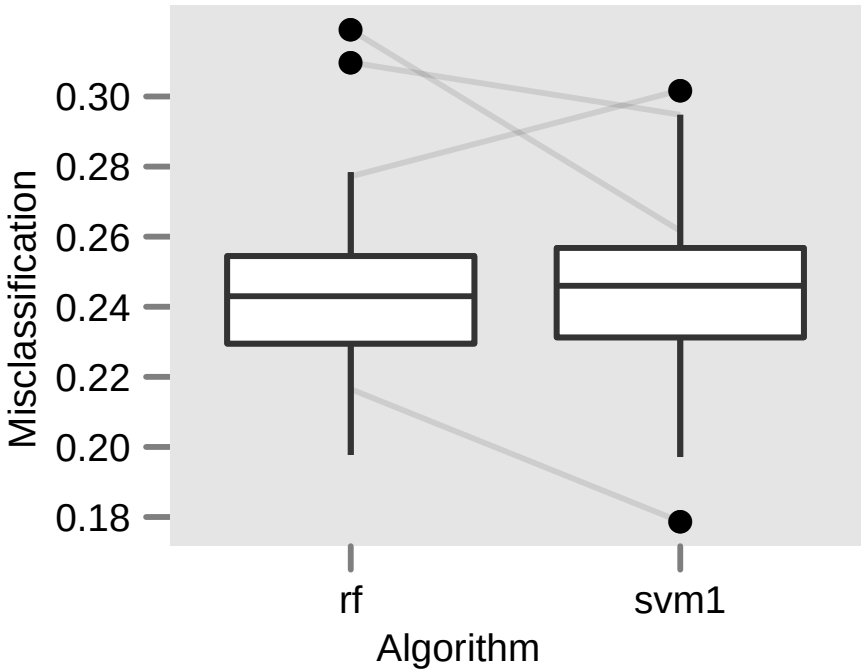
p-value $= 1.312e - 12$

$\Rightarrow$ `lda` $\prec$ `svm1`

Scenario 1 – Different algorithm performances:



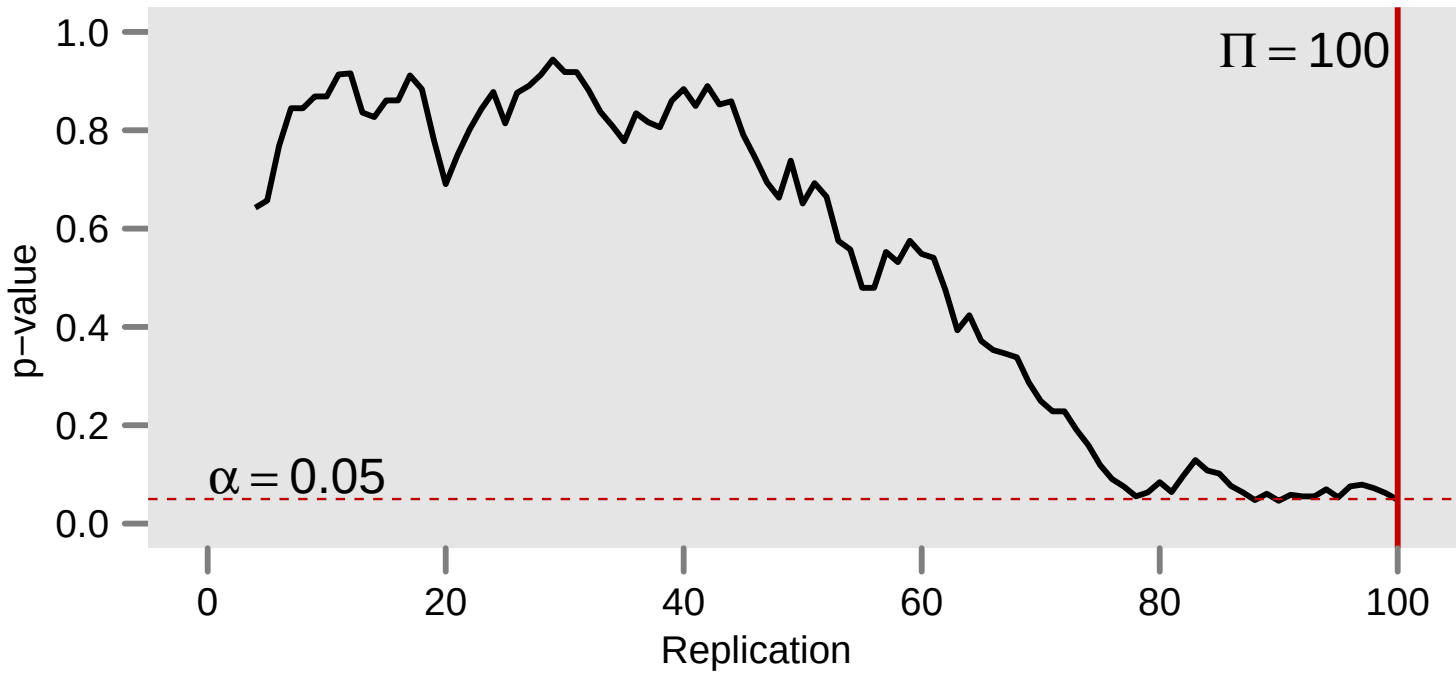
Scenario 2 – Similar algorithm performances:



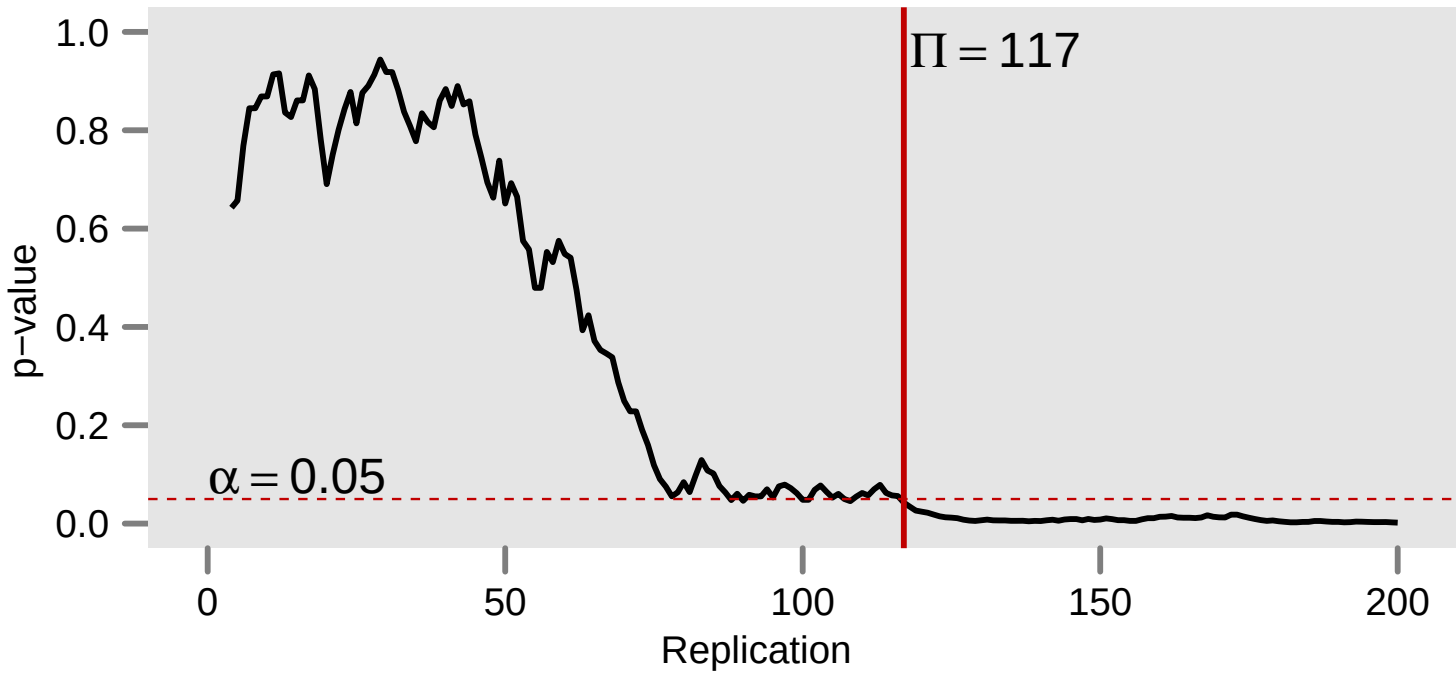
p-value $= 0.0484$

$\Rightarrow$ `rf` $\prec$ `svm1`

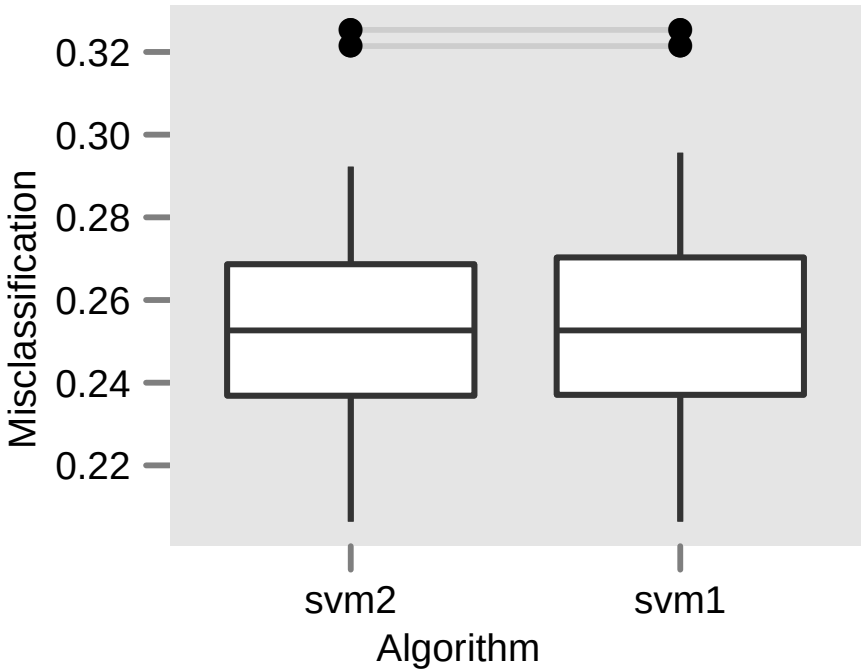
Scenario 2 – Similar algorithm performances:



Scenario 2 – Similar algorithm performances:



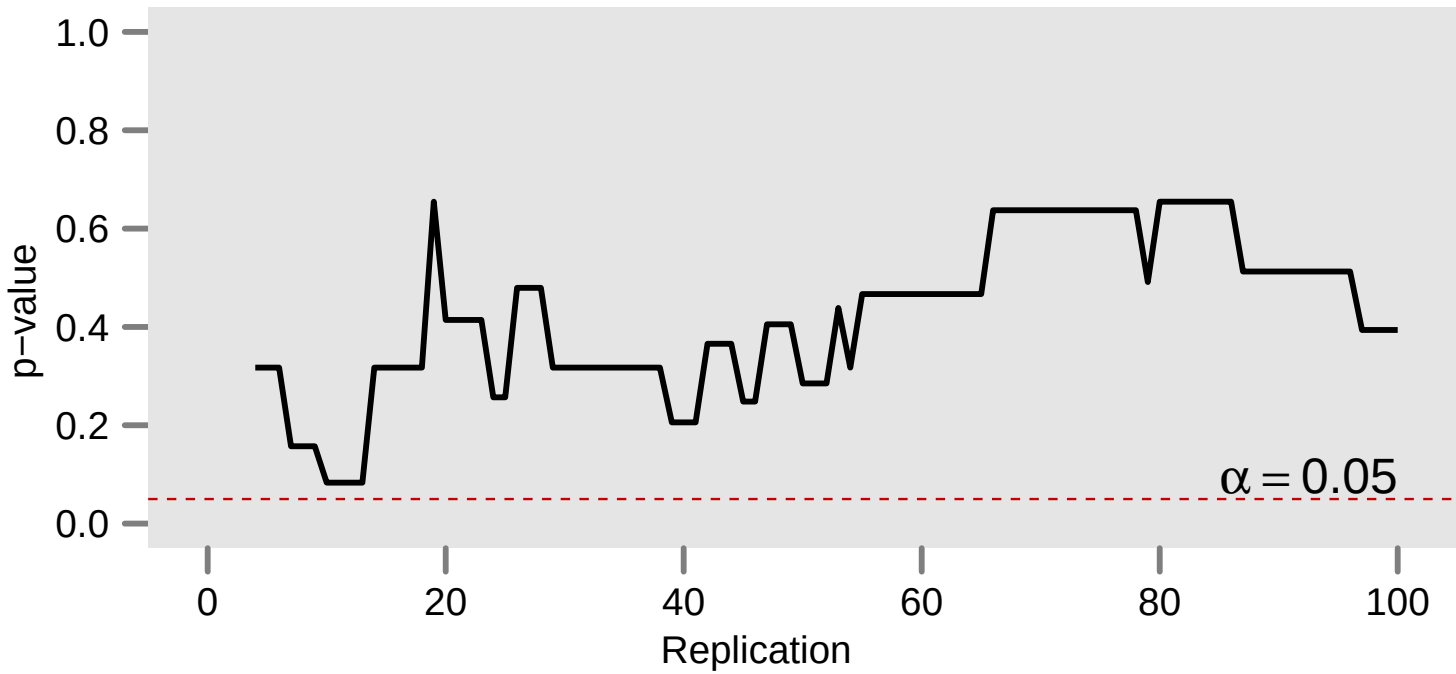
Scenario 3 – Equal algorithm performances:



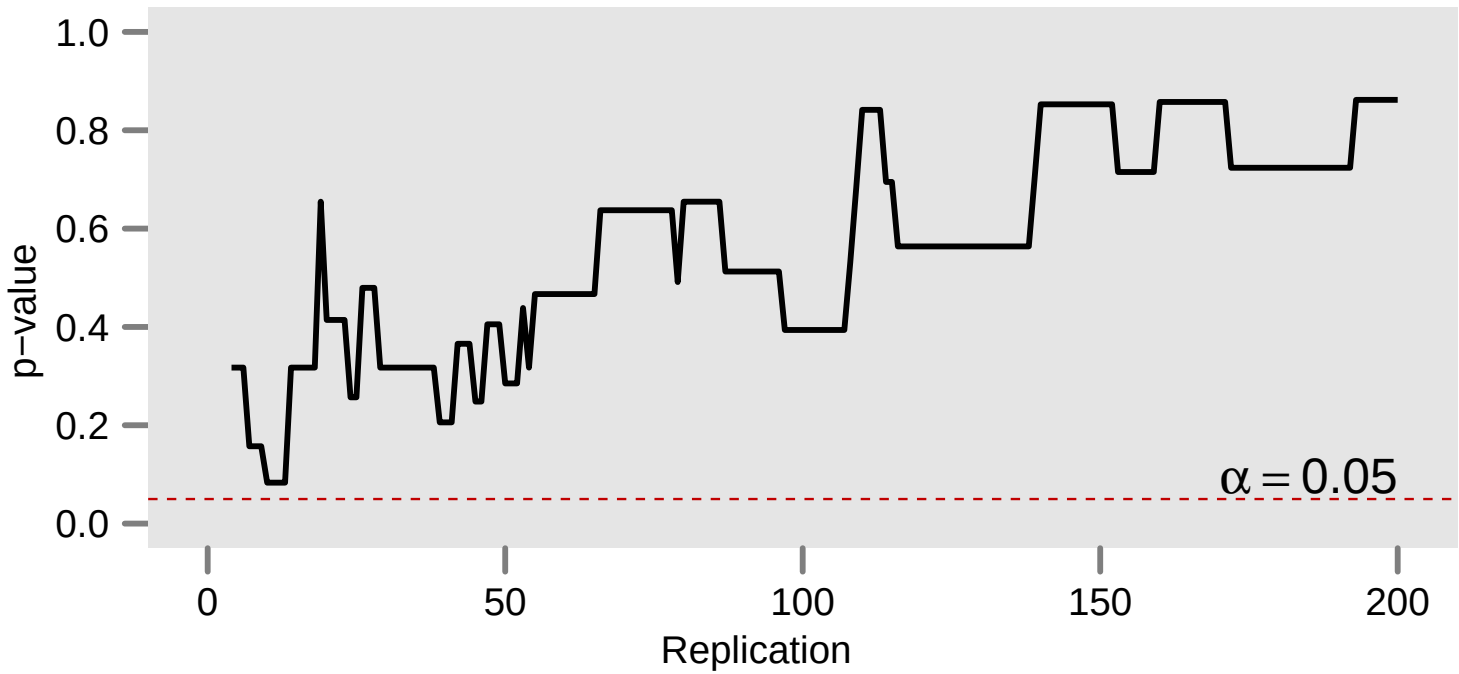
p-pvalue $= 0.9302$

$\Rightarrow$ `svm2` $\approx$ `svm1`

Scenario 3 – Equal algorithm performances:



Scenario 3 – Equal algorithm performances:



Interpretation

Point consecutively significance:

$$\Pi_{\text{Scenario 1}} = 13,\ \Pi_{\text{Scenario 2}} = 117,\ \Pi_{\text{Scenario 3}} = \infty$$

Measure of "how big the difference" is – indicator for relevance?

Decision making

Execute a benchmark experiment as long as needed – either until H_0 is rejected or H_0 is "accepted" (failed to reject).

Repeated significance testing:

Testing not once but multiple times causes the inflation of the probability for the error of the first kind, i.e., the probability of rejecting the global null hypothesis when in fact this hypothesis is true; known as **alpha inflation**.

(*) First addressed by Armitage, McPherson, and Rowe (1969).

Analyses on accumulating data

Sequential: Sample observations one by one; the test is executed after each new observation – the experiment can be stopped at any point.

Group sequential: Sample groups of observations; the test is executed after each group – the experiment can be stopped after each group.

Adaptive: Group sequential with more flexibility, e.g., to change hypothesis, group sample size, etc.

(*) Following Vandemeulebroecke (2008).

Sequential analysis of benchmark experiments:

General differences to the common field (e.g., clinical trials).

1. Compared to clinical trials it is easy and (relatively) cheap to make additional replications until a final decision, i.e., to reject or accept H_0; so, (theoretically) there is no undecidable situation.
2. Benchmark experiments are computer experiments often executed on remote servers, etc; so decisions made in the interim and planning phases need a sound automatization (or "interactive" interim and planning phases).

Recursive combination tests

Adaptive test procedure based on the recursive application of two-stage combination tests; and a p-value function to combine p-values from two stages.

(*) Defined by Brannath, Posch, and Bauer (2002).

Two-stage combination tests:

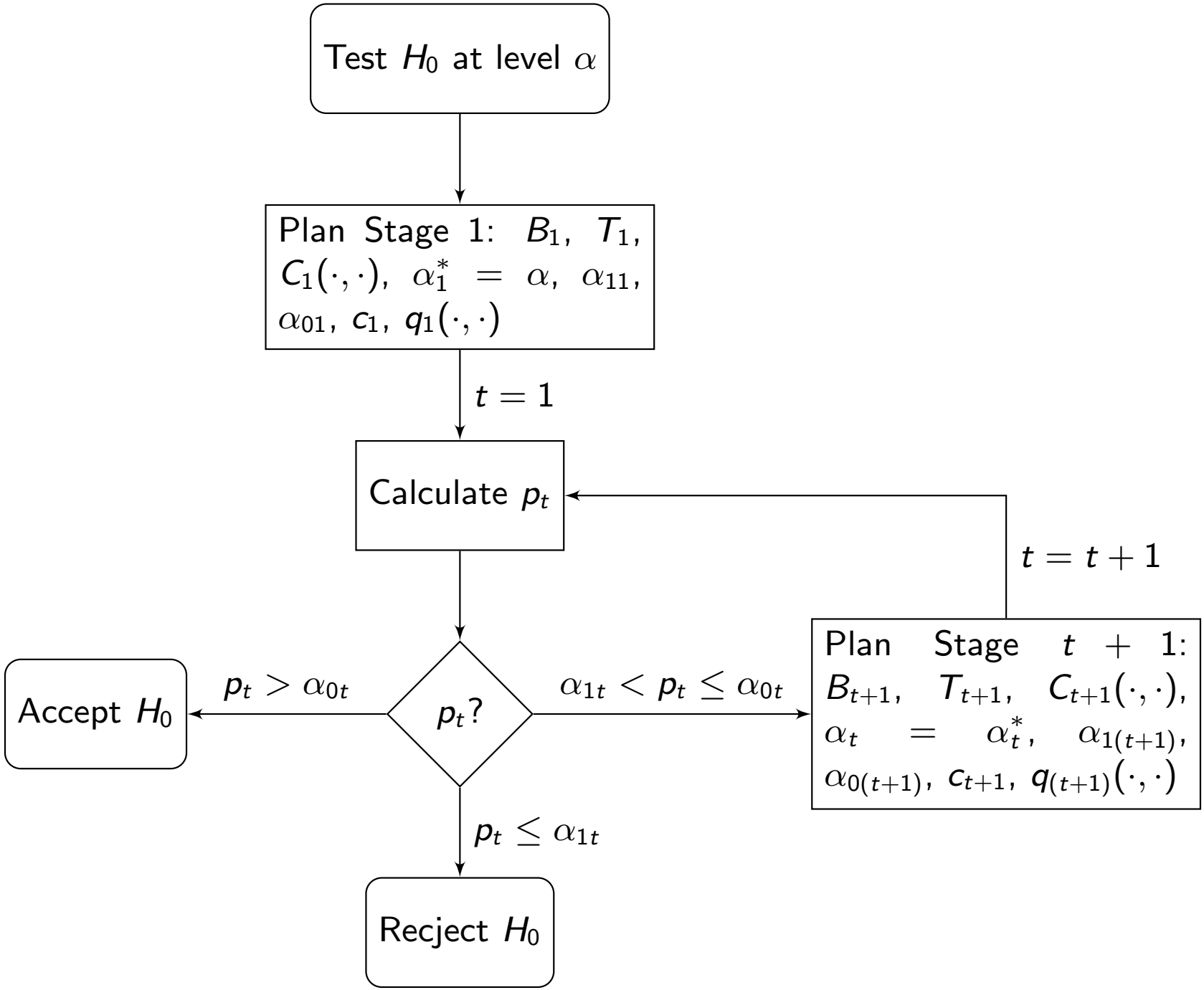
Test a one-sided null hypothesis H_0 at level α using two stages.

Stage 1: B_1 learning samples, decision boundaries α_{01} and α_{11} ($0 \leq \alpha_{11} < \alpha < \alpha_{01} \leq 1$).

$$\text{Decision} = \begin{cases} \text{reject } H_0, & p_1 \leq \alpha_{11} \\ \text{accept } H_0, & p_1 > \alpha_{01} \\ \text{perform Stage 2}, & \alpha_{11} < p_1 \leq \alpha_{01} \end{cases}$$

Stage 2: B_2 learning samples.

$$\text{Decision} = \begin{cases} \text{reject } H_0, & C(p_1, p_2) \leq c \\ \text{undecidable}, & \text{otherwise} \end{cases}$$


Test H_0 at level α
Plan Stage 1: B_1, T_1, $C_1(\cdot,\cdot)$, $\alpha_1^* = \alpha$, α_{11}, α_{01}, c_1, $q_1(\cdot,\cdot)$
$t = 1$
Calculate p_t
$t = t + 1$
Plan Stage $t+1$: B_{t+1}, T_{t+1}, $C_{t+1}(\cdot,\cdot)$, $\alpha_t = \alpha_t^*$, $\alpha_{1(t+1)}$, $\alpha_{0(t+1)}$, c_{t+1}, $q_{(t+1)}(\cdot,\cdot)$
$p_t > \alpha_{0t}$
p_t?
$\alpha_{1t} < p_t \leq \alpha_{0t}$
Accept H_0
$p_t \leq \alpha_{1t}$
Recject H_0

Fisher's combination test:

$$C(p_{t-1}, p_t) = p_{t-1} \cdot p_t$$

Critical value:

$$c_t = \frac{\alpha_t^* - \alpha_{1t}}{\ln \alpha_{0t} - \ln \alpha_{1t}}$$

Conditional significance level:

$$\alpha_1^* = \alpha,\ \alpha_t^* = \frac{c_{t-1}}{p_{t-1}}$$

Conditional decision boundaries:

$$\alpha_{1t} < \alpha_t^* \leq \alpha_{0t}$$

Median combination test:
There is a strong dependence between the p-values, which is caused by the overlap of learning and validation samples in one replication with those of another replication.

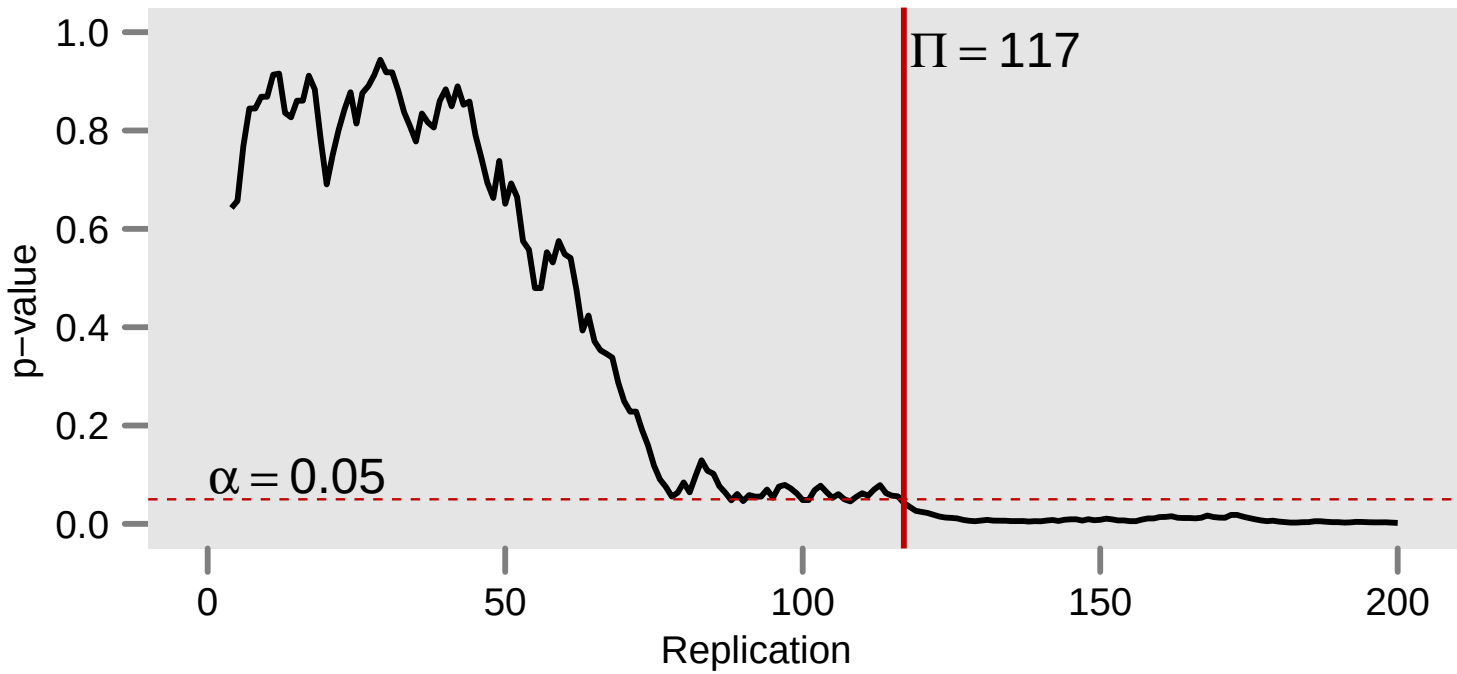
Fisher's combination test is inappropriate here; van de Wiel, Berkhof, and van Wieringen (2009) propose to use

$$C(p_{t-1}, p_t) = median(p_{t-1}, p_t)$$

as p-value combination function.

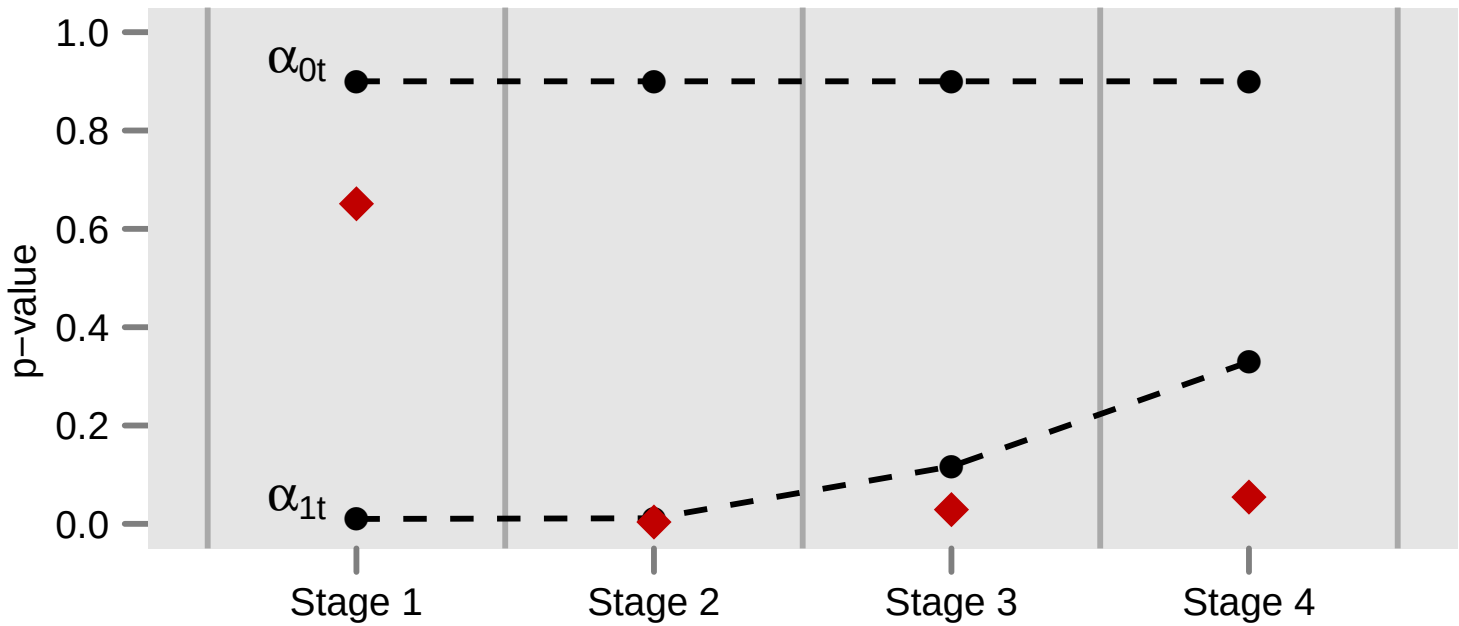
Scenario 2 – Similar algorithm performances:

Split the scenario into four stages, i.e., $B_t = \frac{B}{4} = 50$; define $\alpha = \alpha_1^* = 0.05$, $\alpha_{11} = 0.01$, $\alpha_{01} = 0.9$; and define the rule for the conditional decision boundaries as $\alpha_{0t} = \frac{\alpha_t^*}{1.2}$ and $\alpha_{1t} = \alpha_{1(t-1)}$.



Scenario 2 – Similar algorithm performances:

Split the scenario into four stages, i.e., $B_t = \frac{B}{4} = 50$; define $\alpha = \alpha_1^* = 0.05$, $\alpha_{11} = 0.01$, $\alpha_{01} = 0.9$; and define the rule for the conditional decision boundaries as $\alpha_{0t} = \frac{\alpha_t^*}{1.2}$ and $\alpha_{1t} = \alpha_{1(t-1)}$.



Summary

Monitoring:

- Point of consecutively significance.

Decision making:

- Statistical justification for the number of replications.

- Sound interim phase (or better an “interactive” one)?
- General advantage in real-world benchmark experiments?

Green Benchmarking!

`http://CRAN.R-project.org/package=benchmark`

References

P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated significane test on accumulating data. *Journal of the Royal Statistical Society*, 132(2), 1969.

Werner Brannath, Martin Posch, and Peter Bauer. Recursive combination tests. *Journal of the American Statistical Association*, 97(457):236–244, 2002.

Manuel J. A. Eugster, Torsten Hothorn, and Friedrich Leisch. Exploratory and inferential analysis of benchmark experiments. Under review, 2010a.

Manuel J. A. Eugster, Torsten Hothorn, and Friedrich Leisch. Domain-based benchmark experiments: Exploratory and inferential analysis. Under review, 2010b.

Torsten Hothorn, Friedrich Leisch, Achim Zeileis, and Kurt Hornik. The design and analysis of benchmark experiments. *Journal of Computational and Graphical Statistics*, 14(3):675–699, 2005.

Mark A. van de Wiel, Johannes Berkhof, and Wessel N. van Wieringen. Testing the prediction error difference between 2 predictors. *Biostatistics*, 10(3):550–560, 2009.

Marc Vandemeulebroecke. Group sequential and adaptive designs – a review of basic concepts and points of discussion. *Biometrical Journal*, 50(3), 2008.

Wikipedia. Decathlon. Wikipedia article, visited on November 16, 2010, 2010. URL `http://en.wikipedia.org/wiki/Decathlon`.