The Design and Analysis of Benchmark Experiments – Part II: Analysis

Torsten Hothorn **Achim Zeileis** **Friedrich Leisch** **Kurt Hornik**

Friedrich–Alexander–Universität Erlangen–Nürnberg

http://www.imbe.med.uni-erlangen.de/~hothorn/

Benchmark Experiments

A comparison of algorithms with respect to certain performance measures is of special interest in the following problems

- select the best out of a set of candidates,
- identify groups of algorithms with the same performance,
- test whether any useful structure is inherent in the data or
- demonstrate equivalence of two algorithms.

Illustrating Example

Stabilization of a Linear Discriminant Analysis (LDA) by using low-dimensional Principal Component (PC-q) scores (Läuter, 1992; Läuter et al., 1998; Kropf, 2000) for Glaucoma diagnosis (Hothorn et al., 2003; Mardin et al., 2003).

Laser-scanning images from 98 patients and 98 controls ($n = 196$), $p = 62$ numeric input variables.

Data generating process: The empirical distribution function $\hat{Z}_n$.

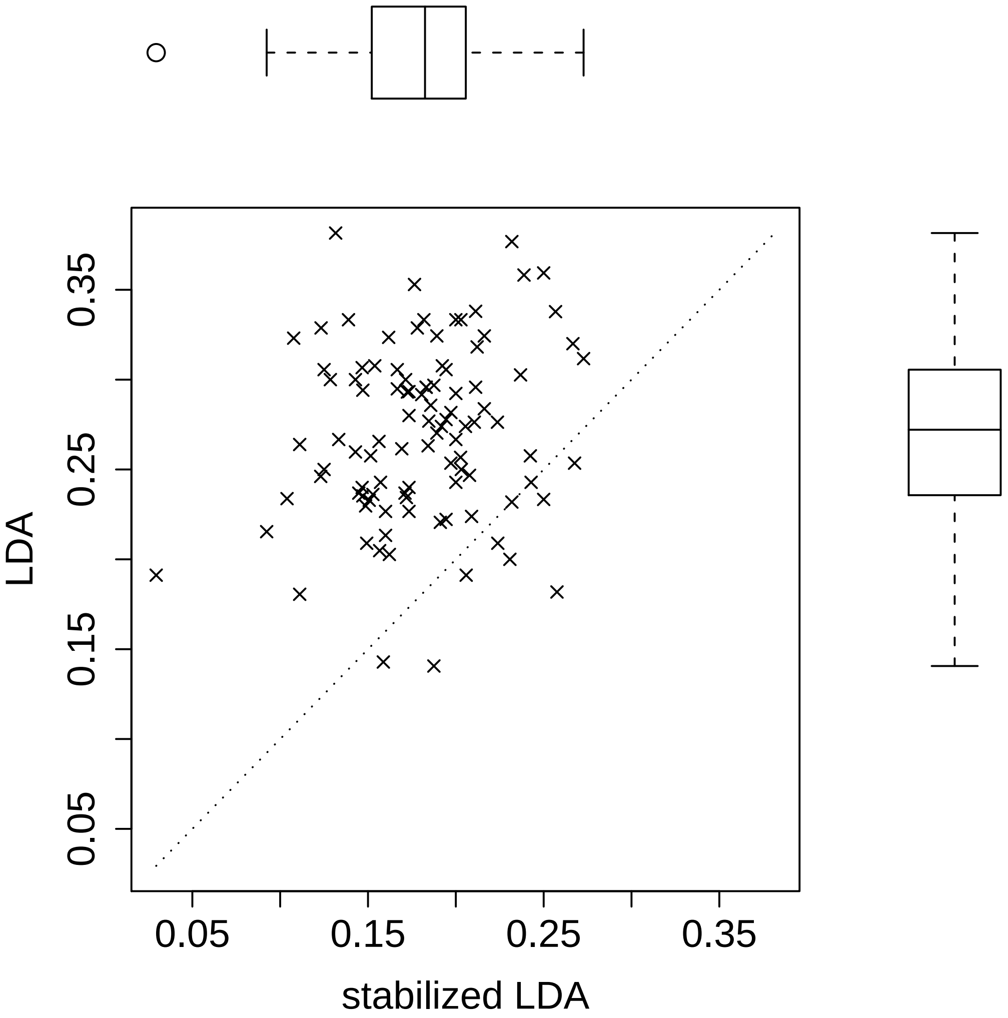
Performance measure: Out-of-bootstrap misclassification error.

Experiment

Question: Does the performance distribution $\hat{P}_{\mathsf{LDA}}(\hat{Z}_n)$ of a LDA using the original p input variables differ from the performance distribution $\hat{P}_{\mathsf{sLDA}}(\hat{Z}_n)$ of a stabilized LDA?

Experiment: Draw B samples $\mathcal{L}^b$ from the data generating process $\hat{Z}_n$ and compute $\hat{p}_{\mathsf{LDA},b}$ and $\hat{p}_{\mathsf{sLDA},b}$, the misclassification errors evaluated on the out-of-bootstrap observations.

Benchmark Experiments



Inference

$$H_0 : \hat{P}_{\mathsf{LDA}}(\hat{Z}_n) = \hat{P}_{\mathsf{sLDA}}(\hat{Z}_n)$$

Problem: We do not know anything about the performances, except that parametric assumptions are surely not appropriate.

Solution: Dispose the performance distributions by conditioning on all permutations of the labels for each bootstrap sample.

Inference

$$T = \sum_{b=1}^{B} \hat{p}_{\textsf{LDA},b} - \hat{p}_{\textsf{sLDA},b} = B(\bar{p}_{\textsf{LDA},\cdot} - \bar{p}_{\textsf{sLDA},\cdot})$$

The conditional distribution of the test statistic T under the conditions described by H_0 can be used to construct a permutation test.

In our case, the P-value based on the asymptotic conditional distribution is $p < 0.001$ and therefore H_0 can be rejected.

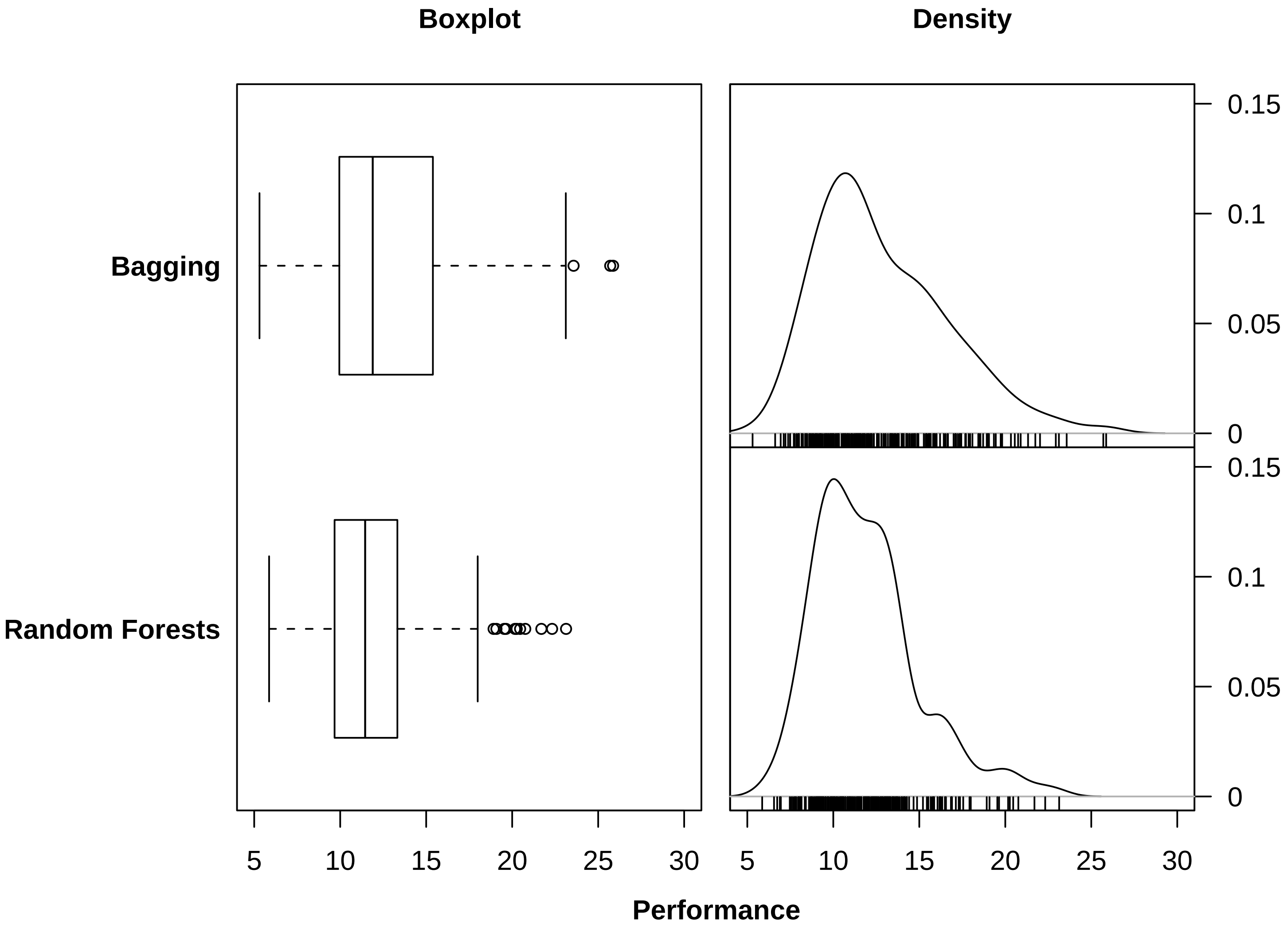
A Regression Example

Exactly the same methodology can be applied to regression problems with univariate numeric responses. Example: Can additional randomness via Random Forests improve Bagging for the Boston Housing data?

House prices for $n = 506$ houses near Boston, $p = 13$ input variables.

Data generating process: The empirical distribution function $\hat{Z}_n$.

Performance measure: Out-of-bootstrap mean squared error.


Boxplot
Density
Bagging
Random Forests
5
10
15
20
25
30
0.15
0.1
0.05
0
Performance

Inference

The null-hypothesis of equal performance distributions can be rejected (P-value < 0.001).

The estimated difference of the mean square error of Bagging compared to Random Forests is 0.969 with confidence limits $(0.633, 1.305)$.

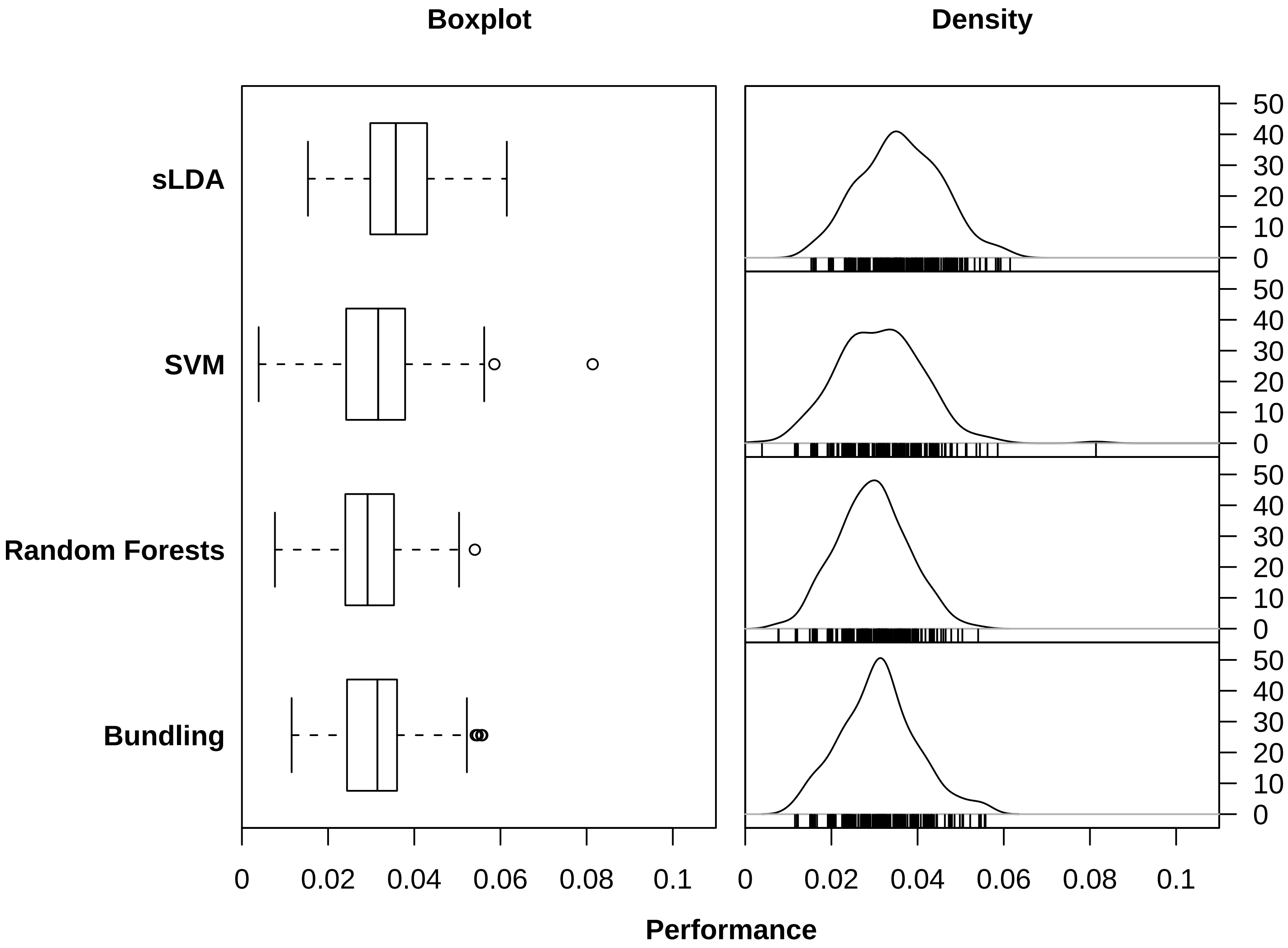
Comparison of Multiple Algorithms

When multiple algorithms are under test, we are interested in both a global test and a multiple test procedure showing where the differences, if any, come from. Example: Breast Cancer data with tumor classification from $n = 699$ observations with $p = 9$ inputs.

Comparison of sLDA, Support Vector Machine, Random Forests and Bundling (Hothorn and Lausen, 2003).

Data generating process: The empirical distribution function $\hat{Z}_n$.

Performance measure: Out-of-bootstrap misclassification error.


Boxplot
Density
sLDA
SVM
Random Forests
Bundling
0
0.02
0.04
0.06
0.08
0.1
0
10
20
30
40
50
Performance

Inference

Again, the global hypothesis

$$H_0 : \hat{P}_1(\hat{Z}_n) = \ldots = \hat{P}_K(\hat{Z}_n)$$

can be rejected (P-value < 0.001).

Problem: Which differences 'cause' the rejection of H_0?

Solution: One can avoid complicated closed testing procedures by computing confidence intervals after mapping the B-block design into a K-sample problem via alignment (Hájek et al., 1999).

Alignment

When we look at the performance measure of algorithm k in the bth sample drawn from the data generating process, we might want to write

$$p_{kb} = \mu + \beta_b + \gamma_k + \varepsilon_{kb}$$

where μ corresponds to the performance of the Bayes-rule, β_b is the error induced by the b sample and γ_k is the error of the kth algorithm, the quantity we are primarily interested in, ε indicates an error term.

Alignment (cont'd)

The aligned performance measures $p^\star_{kb}$ cover the difference of the performance of the kth algorithm from the average performance of all K algorithms:

$$p^\star_{kb} = p_{kb} - \bar{p}_{\cdot b} = (\gamma_k + \varepsilon_{kb}) - \frac{1}{K}\sum_{k=1}^{K}(\gamma_k + \varepsilon_{kb})$$

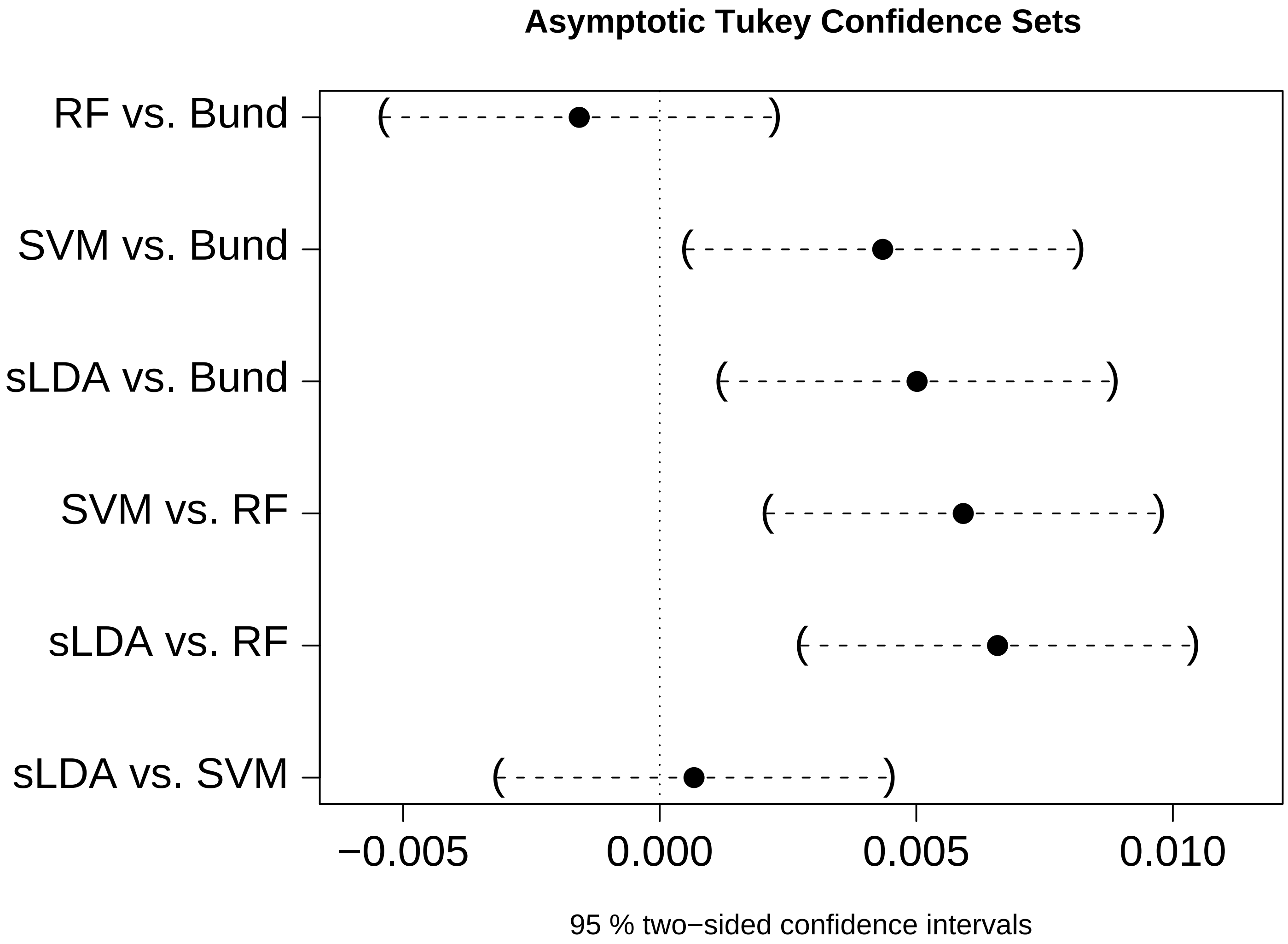
For classification problems, $p^\star_{k_1 b} - p^\star_{k_2 b}$ is the difference of the misclassification error.

Alignment (cont'd)

The aligned random variables are not independent but exchangeable for each of the b samples and are independent between samples.

Therefore, (asymptotic) permutation test procedures can be used to assess the deviations from the global null-hypothesis.

For example, asymptotic simultaneous confidence intervals for Tukey-contrasts can be used for an all-pair comparison of the K algorithms under test.


Asymptotic Tukey Confidence Sets
RF vs. Bund
SVM vs. Bund
sLDA vs. Bund
SVM vs. RF
sLDA vs. RF
sLDA vs. SVM
−0.005
0.000
0.005
0.010
95 % two−sided confidence intervals

Classical Tests?

We advocate usage of permutation tests, but what about more classical tests?

Consider a paired comparison of sLDA vs. SVM for the Breast Cancer data:

- Permutation test: $T = 1.488, \quad p = 0.776$
- t test: $t = 0.284, \quad p = 0.777$
- Wilcoxon signed rank test: $W = 18216, \quad p < 0.001$

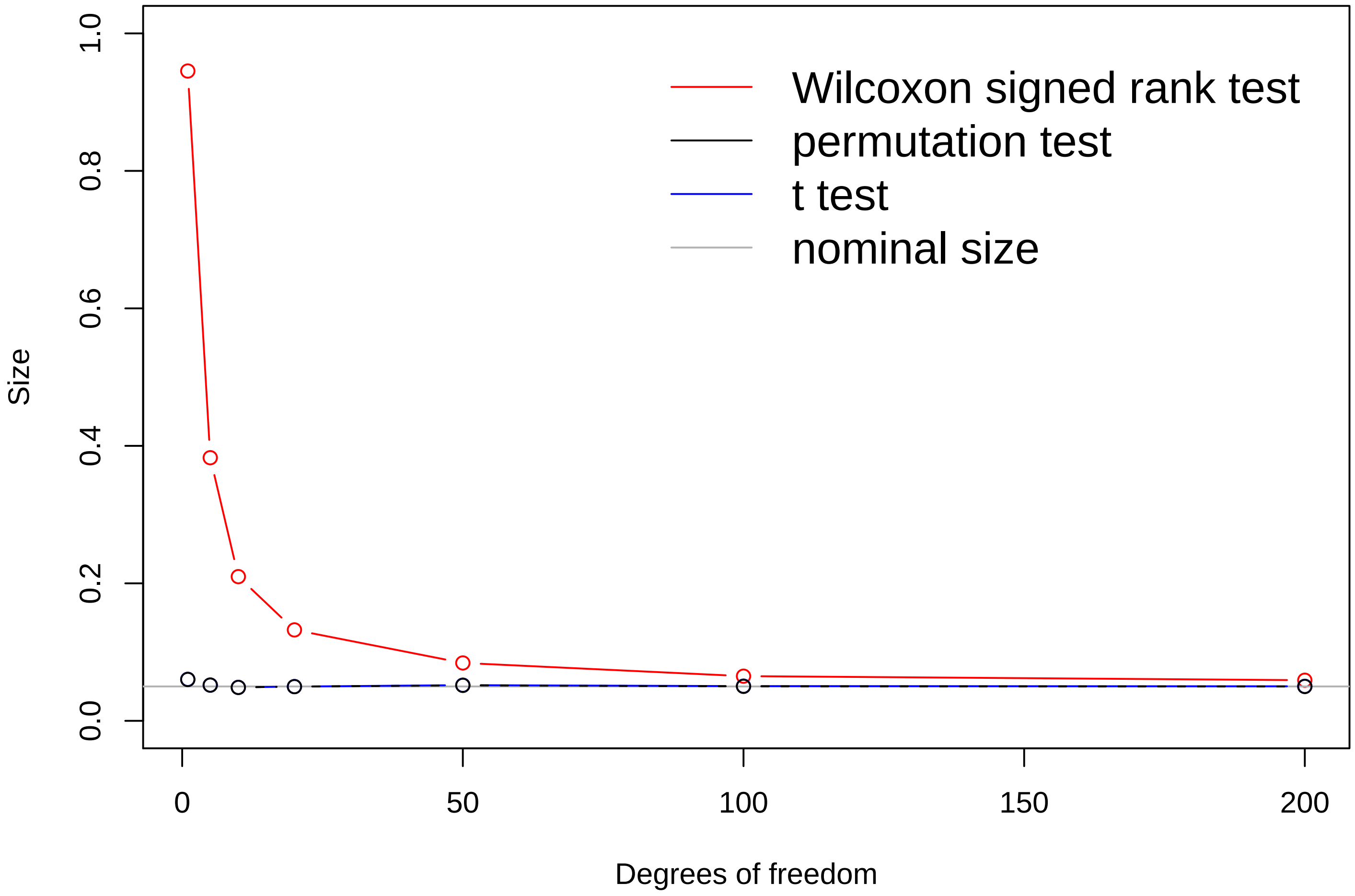
Rank Tests: A Warning

Tests like the Wilcoxon signed rank test are constructed for the null-hypothesis ‘the difference of the performance measures is **symmetrically** distributed around zero’. For non-symmetric distributions this leads to a complete desaster.

Look at $n = 500$ realizations of a skewed random variable

$$\frac{X - d}{\sqrt{2d}}$$

with expectation zero and unit variance with $X \sim \chi^2_d$.

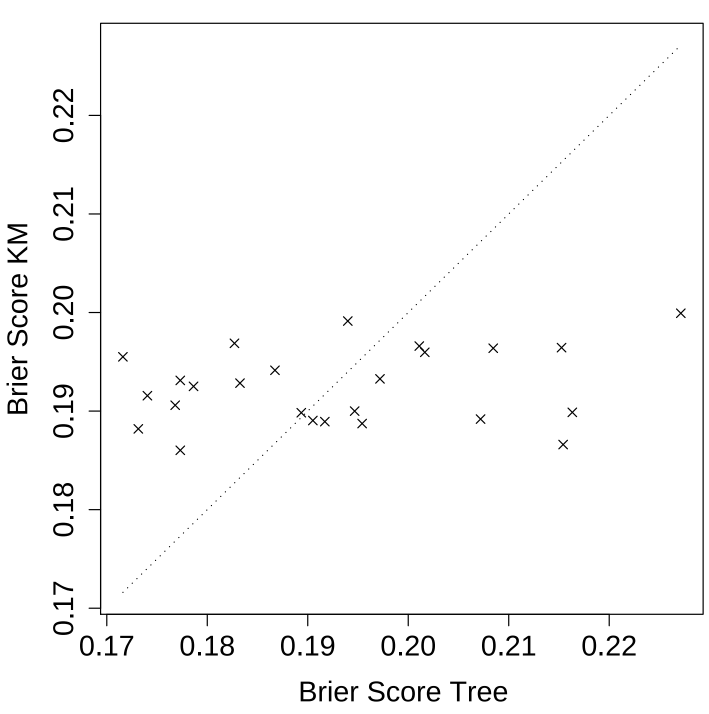

Wilcoxon signed rank test
permutation test
t test
nominal size
Size
0.0
0.2
0.4
0.6
0.8
1.0
Degrees of freedom
0
50
100
150
200

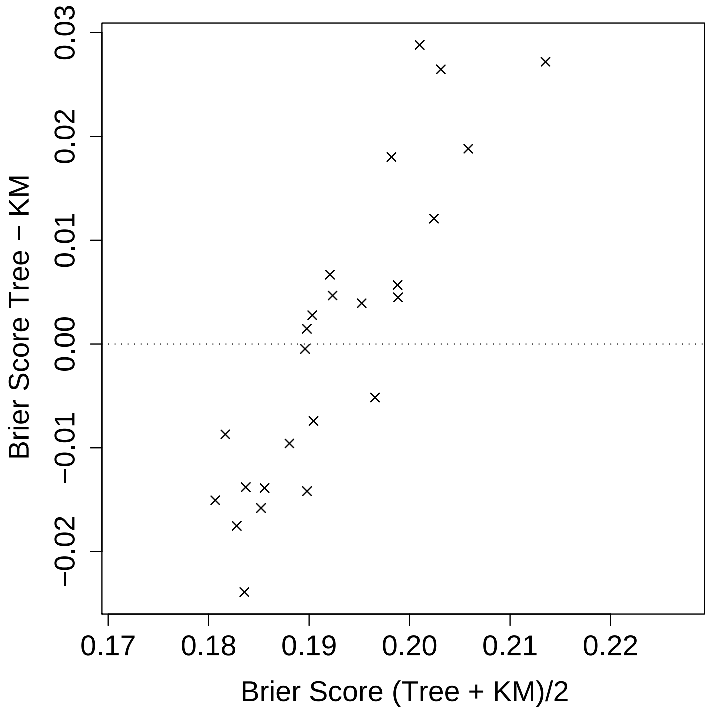
Lifetime Analysis Problems

Appropriate performance measures for censored responses are by no means obvious and still a matter of debate (Henderson, 1995; Graf et al., 1999; Molinaro et al., 2004). We use the Brier score for censored data suggested by Graf et al. (1999).

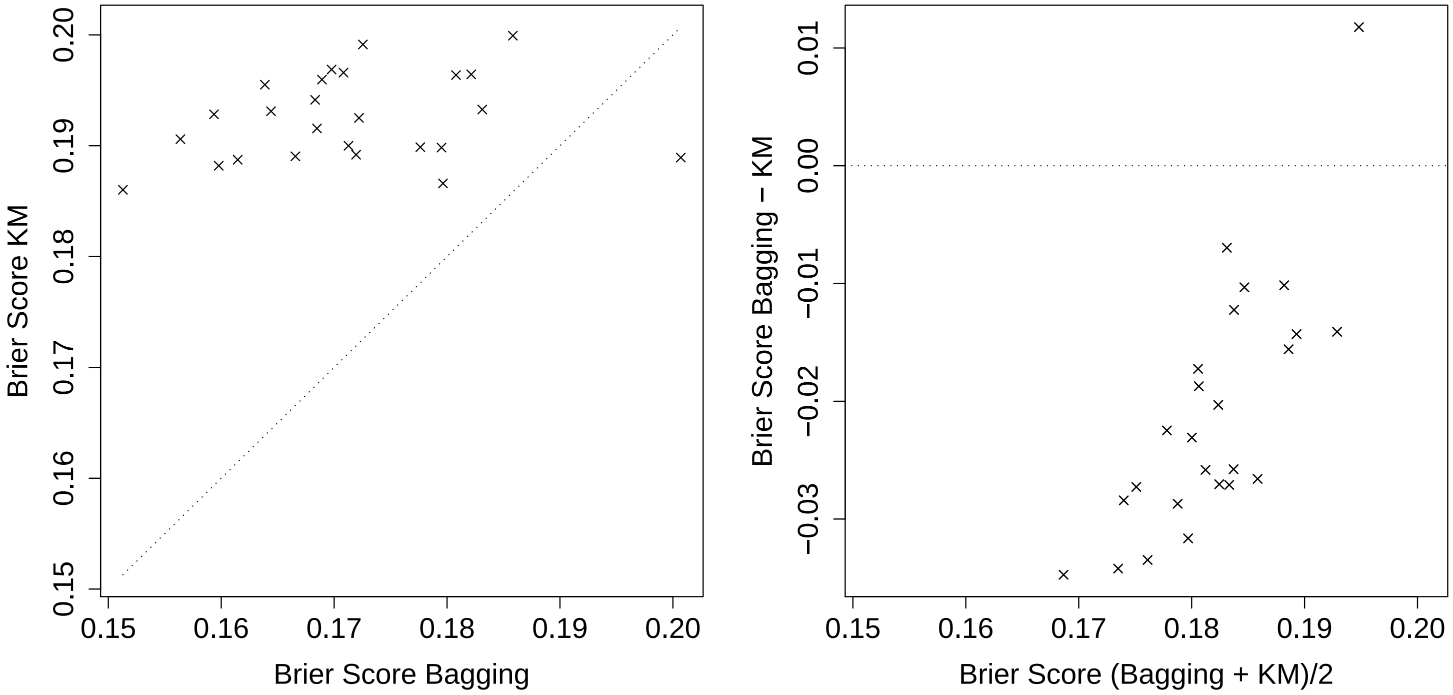
Example: Predictive performance of the Kaplan-Meier estimator, a single survival tree and Bagging of survival trees (Hothorn et al., 2004) measured for $n = 686$ women enrolled in the German Breast Cancer Study (Group 2).

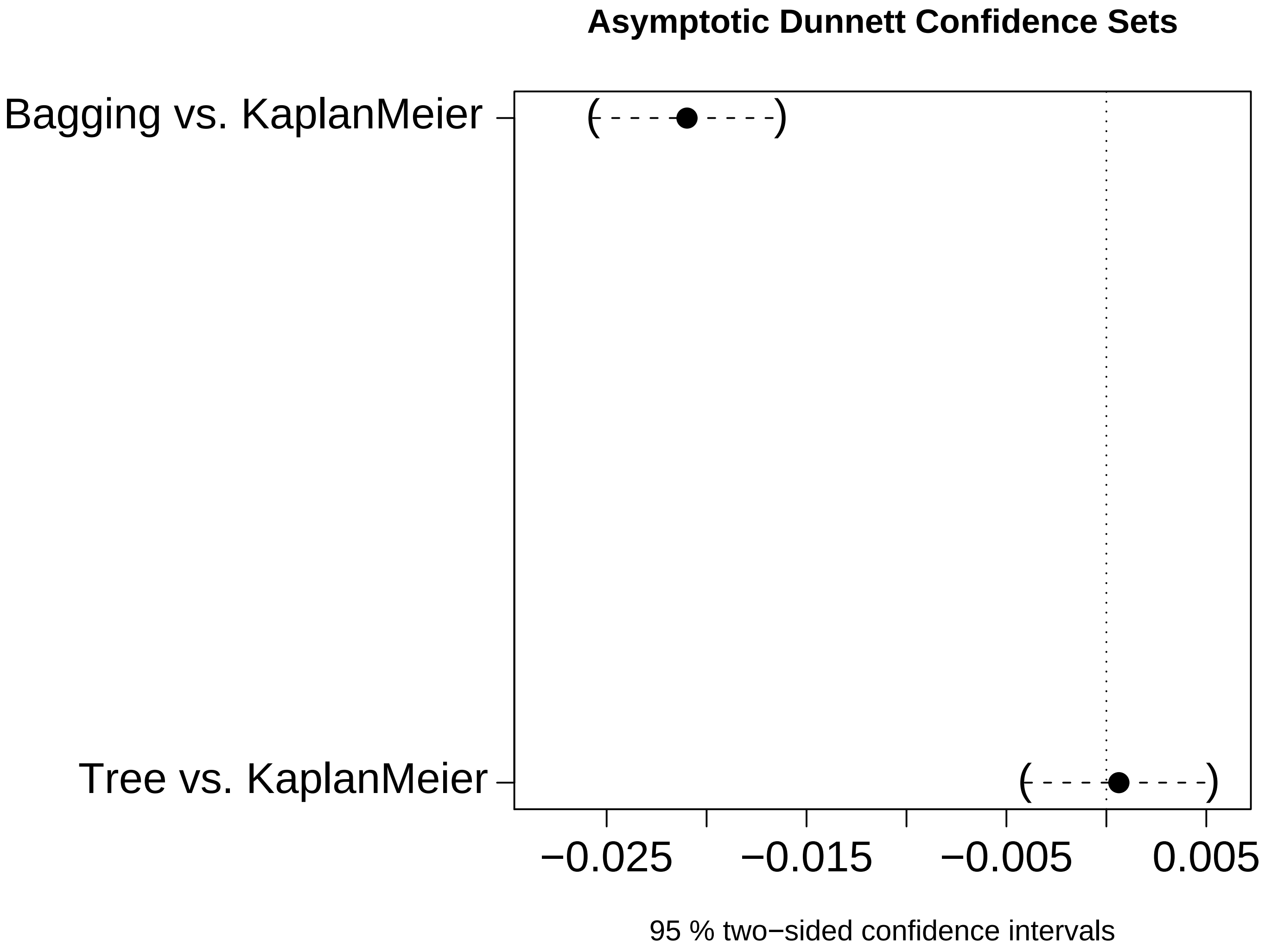
Kaplan-Meier vs. Single Tree





Kaplan-Meier vs. Bagging




Asymptotic Dunnett Confidence Sets
Bagging vs. KaplanMeier
Tree vs. KaplanMeier
−0.025
−0.015
−0.005
0.005
95 % two−sided confidence intervals

Interpretation

Predictions derived from the estimated Kaplan-Meier curve don't take any information covered by the input variables into account. A test for the hypothesis

> *there is no (detectable) relationship between the input variables and the response*

can therefore be performed by comparing the performance of the simple Kaplan-Meier curve with the performance of the best tools available for predicting survival times.

Conclusion

When comparing the performance of K algorithms it is appropriate to treat the B samples from the data generating process as blocks.

Standard statistical test procedures can be used to compare arbitrary performance measures for multiple algorithms.

Some classical parametric and non-parametric procedures are only sub-optimal, we advocate procedures based on the conditional distribution of test statistics for inference.

References

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999), "Assessment and comparison of prognostic classification schemes for survival data," *Statistics in Medicine*, 18, 2529–2545.

Hájek, J., Šidák, Z., and Sen, P. K. (1999), *Theory of Rank Tests*, London: Academic Press, 2nd edition.

Henderson, R. (1995), "Problems and prediction in survival-data analysis," *Statistics in Medicine*, 14, 161–184.

Hothorn, T. and Lausen, B. (2003), "Bundling classifiers by bagging trees," *Preprint, Friedrich-Alexander-University Erlangen-Nuremberg*, URL http://www.mathpreprints.com/.

Hothorn, T., Lausen, B., Benner, A., and Radespiel-Tröger, M. (2004), "Bagging survival trees," *Statistics in Medicine*, 23, 77–91.

Hothorn, T., Pal, I., Gefeller, O., Lausen, B., Michelson, G., and Paulus, D. (2003), "Automated classification of optic nerve head topography images for glaucoma screening," in *Studies in Classification, Data Analysis, and Knowledge Organization: Exploratory Data Analysis in Empirical Research*, eds. M. Schwaiger and O. Opitz, Heidelberg: Springer, pp. 346–356.

Kropf, S. (2000), *Hochdimensionale multivariate Verfahren in der medizinischen Statistik*, Aachen: Shaker Verlag.

Läuter, J. (1992), *Stabile multivariate Verfahren: Diskriminanzanalyse - Regressionsanalyse - Faktoranalyse*, Berlin: Akademie Verlag.

Läuter, J., Glimm, E., and Kropf, S. (1998), "Multivariate tests based on left-spherically distributed linear scores," *The Annals of Statistics*, 26, 1972–1988, correction: 1999, Vol. 27, p. 1441.

Mardin, C. Y., Hothorn, T., Peters, A., Jünemann, A. G., Nguyen, N. X., and Lausen, B. (2003), "New glaucoma classification method based on standard HRT parameters by bagging classification trees," *Journal of Glaucoma*, 12, 340–346.

Molinaro, A. M., Dudoit, S., and van der Laan, M. J. (2004), "Tree-based multivariate regression and density estimation with right-censored data," *Journal of Multivariate Analysis*, 90, 154–177.