Testing Variable Importance in Random Forests

Carolin Strobl (LMU München) and Achim Zeileis (WU Wien)

lifestat 2008

Introduction

random forests

Introduction

random forests

- have become increasingly popular in, e.g., genetics and the neurosciences

Introduction

random forests

- have become increasingly popular in, e.g., genetics and the neurosciences [imagine a long list of references here]

Introduction

random forests

- have become increasingly popular in, e.g., genetics and the neurosciences [imagine a long list of references here]
- can deal with "small n large p"-problems, high-order interactions, correlated predictor variables

Introduction

random forests

- have become increasingly popular in, e.g., genetics and the neurosciences [imagine a long list of references here]
- can deal with "small n large p"-problems, high-order interactions, correlated predictor variables
- are used not only for prediction, but also to assess variable importance and

Introduction

random forests

- have become increasingly popular in, e.g., genetics and the neurosciences [imagine a long list of references here]
- can deal with "small n large p"-problems, high-order interactions, correlated predictor variables
- are used not only for prediction, but also to assess variable importance and
- on the official random forest website Breiman and Cutler (2008) even suggest a significance test for the variable importance...

The permutation importance

The suggested test

Investigating the statistical properties

Specifying the null hypothesis

Summary and outlook

Random forests

learn a random forest as a classification/regression model

to predict Y from $X_1, \ldots, X_p$

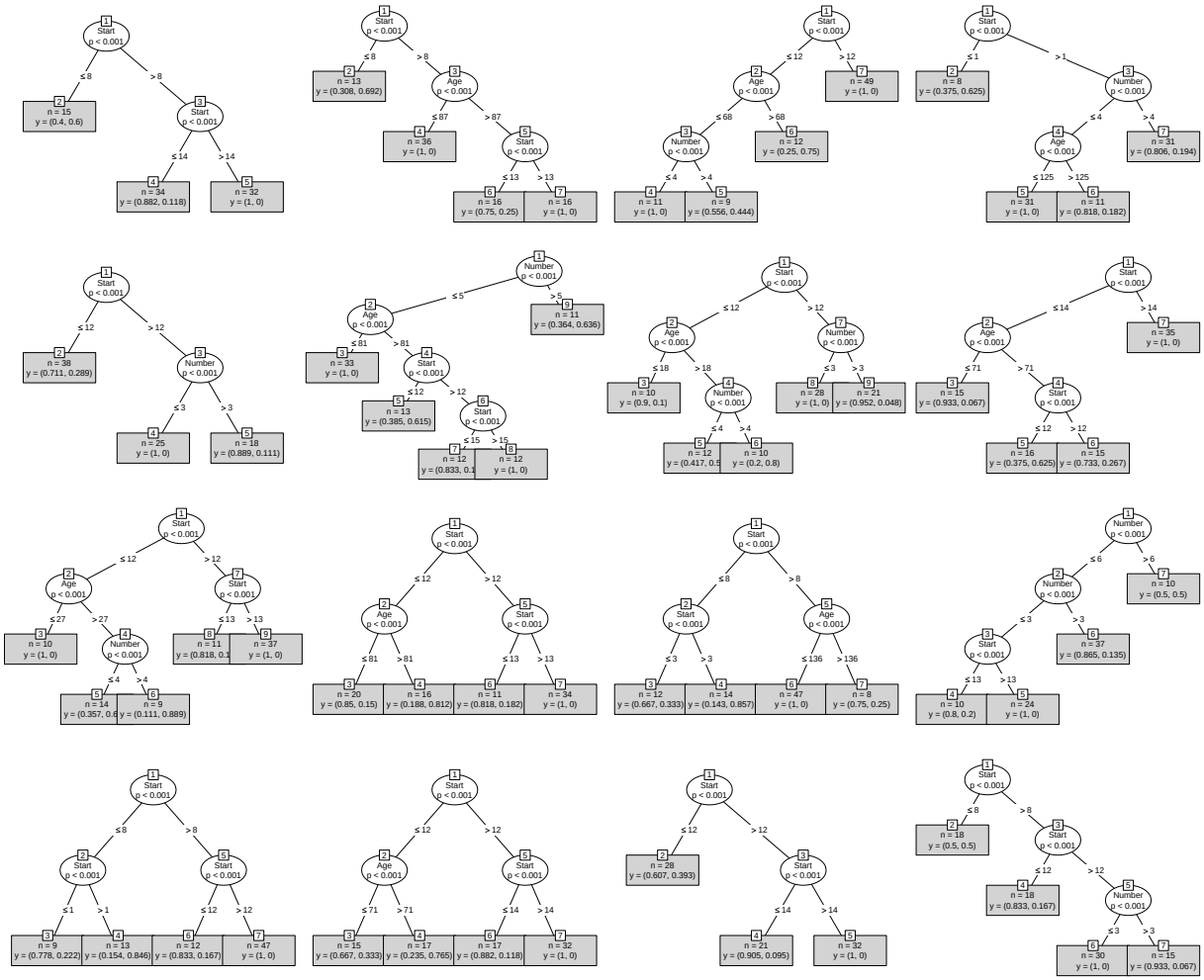
Random forests

learn a random forest as a classification/regression model

to predict Y from $X_1, \ldots, X_p$

result: almost a black-box

(Small) random forest



Measuring variable importance

Measuring variable importance

- Gini importance

  mean Gini gain produced by X_j over all trees

  (can be severely biased due to estimation bias and mutiple testing; Strobl et al., 2007)

Measuring variable importance

- Gini importance

  mean Gini gain produced by X_j over all trees

  (can be severely biased due to estimation bias and mutiple testing; Strobl et al., 2007)

- permutation importance

  mean decrease in classification accuracy after permuting X_j over all trees

  - informative variables produce a systematic decrease in accuracy when permuted
  - uninformative variables produce a random decrease or increase in accuracy when permuted

  (unbiased when subsampling is used; Strobl et al., 2007)

The permutation importance

within each tree t

$$VI^{(t)}(\mathbf{x}_j) = \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_i^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|} - \frac{\sum_{i \in \overline{\mathfrak{B}}^{(t)}} I\left(y_i = \hat{y}_{i,\pi_j}^{(t)}\right)}{\left|\overline{\mathfrak{B}}^{(t)}\right|}$$

$\hat{y}_i^{(t)} = f^{(t)}(\mathbf{x}_i)$ = predicted class before permuting

$\hat{y}_{i,\pi_j}^{(t)} = f^{(t)}(\mathbf{x}_{i,\pi_j})$ = predicted class after permuting X_j

$\mathbf{x}_{i,\pi_j} = \left(x_{i,1}, \ldots, x_{i,j-1}, \; x_{\pi_j(i),j}, x_{i,j+1}, \ldots, x_{i,p}\right)$

Note: $VI^{(t)}(\mathbf{x}_j) = 0$ by definition, if X_j is not in tree t

The permutation importance

over all trees:

1. raw importance

$$VI(\mathbf{x}_j) = \frac{\sum_{t=1}^{ntree} VI^{(t)}(\mathbf{x}_j)}{ntree}$$

The permutation importance

over all trees:

2. scaled importance (z-score)

$$\frac{VI(\mathbf{x}_j)}{\frac{\hat{\sigma}}{\sqrt{ntree}}} = z_j$$

- individual $VI^{(t)}(\mathbf{x}_j)$ have standard deviation σ and

- individual $VI^{(t)}(\mathbf{x}_j)$ have standard deviation σ and
- are computed from *ntree* independent bootstrap samples (where *ntree* is large)

- individual $VI^{(t)}(\mathbf{x}_j)$ have standard deviation σ and
- are computed from *ntree* independent bootstrap samples (where *ntree* is large)
- central limit theorem for the mean $VI(\mathbf{x}_j)$
  $\Rightarrow$ normal with standard error $\sigma/\sqrt{ntree}$

- individual $VI^{(t)}(\mathbf{x}_j)$ have standard deviation σ and
- are computed from *ntree* independent bootstrap samples (where *ntree* is large)
- central limit theorem for the mean $VI(\mathbf{x}_j)$
  $\Rightarrow$ normal with standard error $\sigma/\sqrt{ntree}$

under the null hypothesis of zero importance:

$$z_j \overset{as.}{\sim} N(0,1)$$

The suggested test

if z_j exceeds the α-quantile of N(0,1) $\Rightarrow$ reject the null hypothesis of zero importance for variable X_j

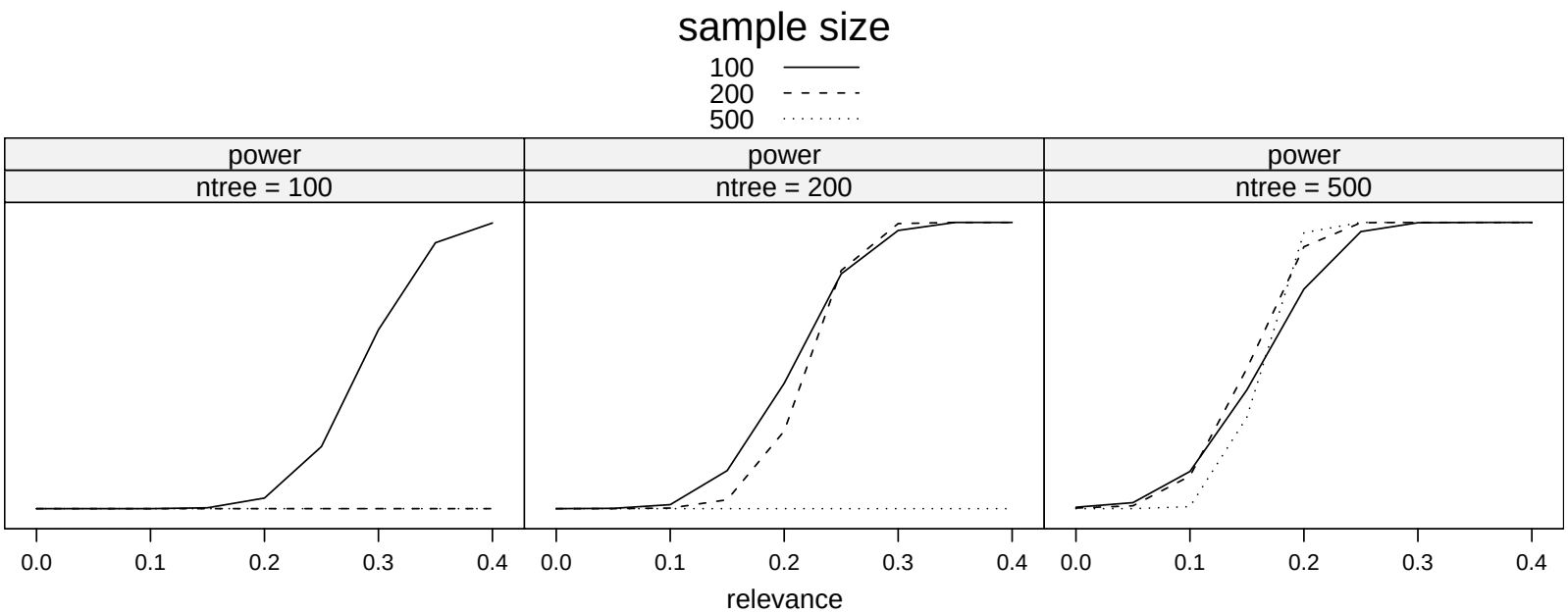
Simulation study

- generate data sets of sample size $n = 100, 200$ and 500
- five predictor variables of which only X_1 is relevant with
- $$y \sim \begin{cases} B(n, 0.5 - \rho) & \text{for} \quad X_1 = 0 \\ B(n, 0.5 + \rho) & \text{for} \quad X_1 = 1 \end{cases}$$ with relevance

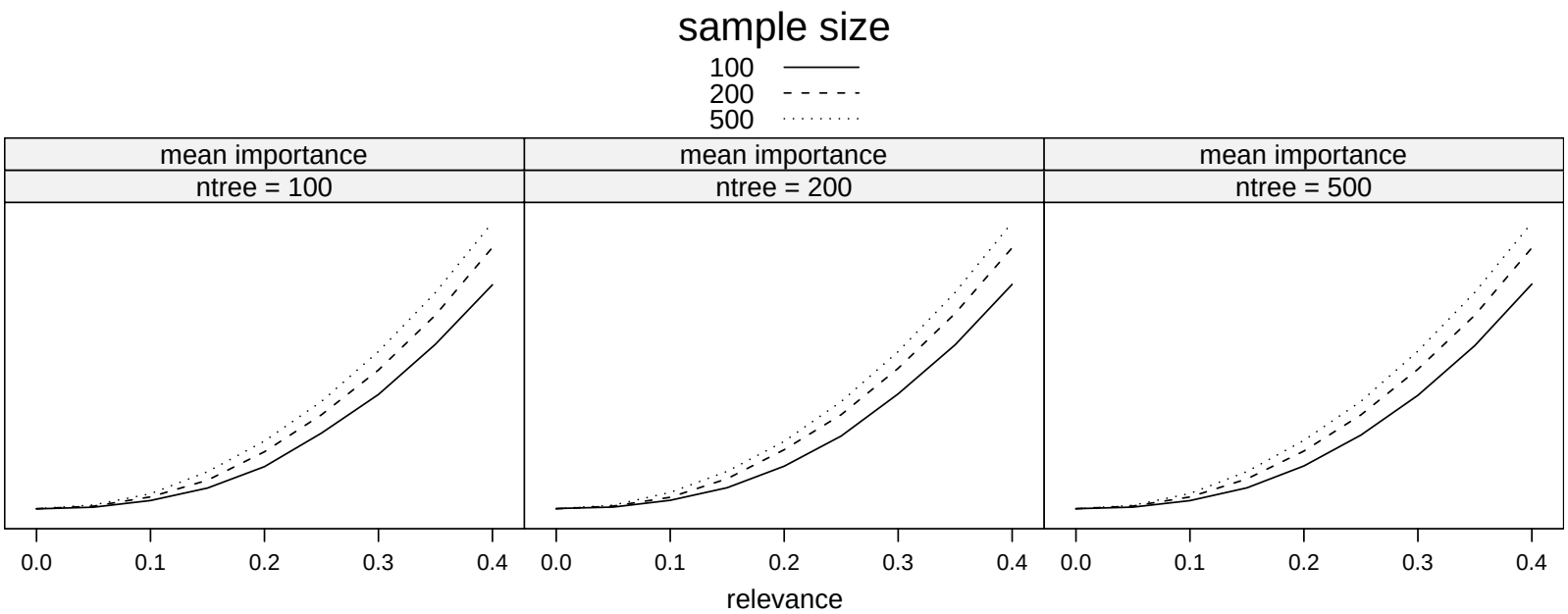
  $\rho = 0, 0.05, \ldots, 0.5$
- fit random forests with *ntree* $= 100, 200$ and 500
- for 1000 iterations in each combination:

  count how many times the null hypothesis for X_1 was rejected

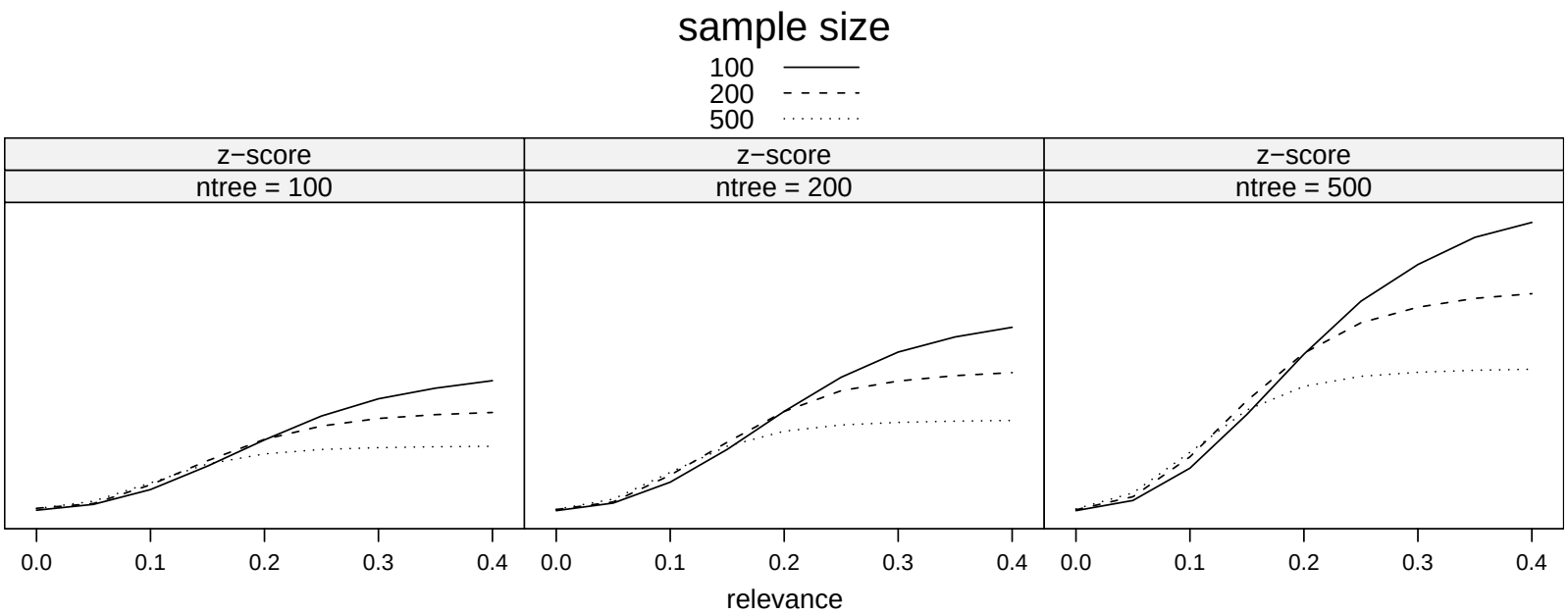
The power

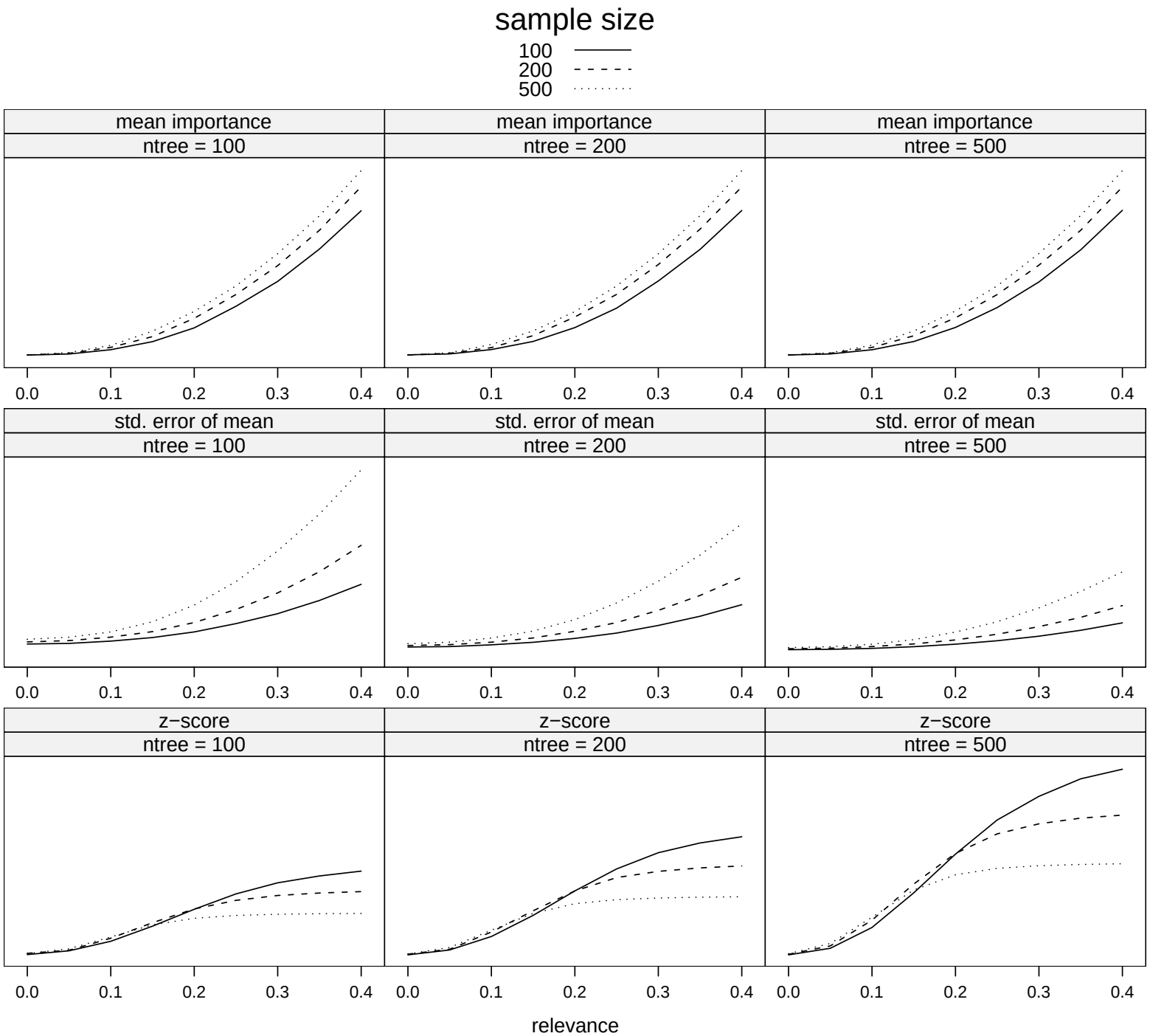


The average raw importance

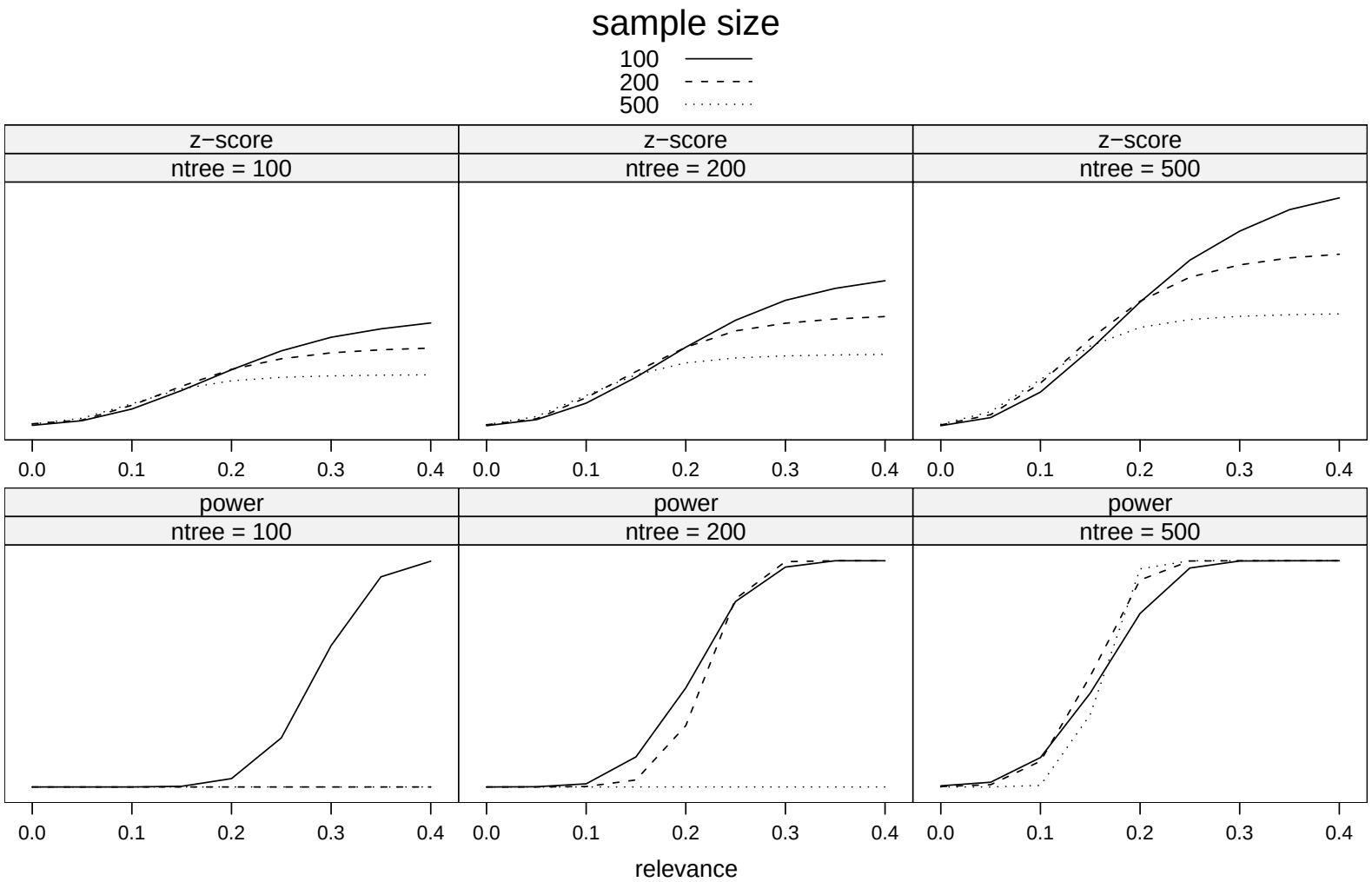


The average z-score





The average z-score and the power



Findings

z-score and power

- increase in the number of trees
- decrease in the sample size

What null hypothesis were we testing in the first place?

obs	Y	X_j	Z
1	y_1	$x_{\pi_j(1),j}$	z_1
$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	y_1	$x_{\pi_j(i),j}$	z_i
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	y_1	$x_{\pi_j(n),j}$	z_n

$$H_0 : X_j \perp Y, Z \text{ or } X_j \perp Y \wedge X_j \perp Z$$

$$P(Y, X_j, Z) \overset{H_0}{=} P(Y, Z) \cdot P(X_j)$$

What null hypothesis were we testing in the first place?

the current null hypothesis reflects independence of X_j from both Y and the remaining predictor variables Z

What null hypothesis were we testing in the first place?

the current null hypothesis reflects independence of X_j from both Y and the remaining predictor variables Z

$\Rightarrow$ a high variable importance can result from violation of either one

Conditional permutation scheme

obs	Y	X_j	Z
1	y_1	$x_{\pi_{j\|Z=a}(1),j}$	$z_1 = a$
3	y_3	$x_{\pi_{j\|Z=a}(3),j}$	$z_3 = a$
27	y_{27}	$x_{\pi_{j\|Z=a}(27),j}$	$z_{27} = a$
6	y_6	$x_{\pi_{j\|Z=b}(6),j}$	$z_6 = b$
14	y_{14}	$x_{\pi_{j\|Z=b}(14),j}$	$z_{14} = b$
33	y_{33}	$x_{\pi_{j\|Z=b}(33),j}$	$z_{33} = b$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

$$H_0 : X_j \perp Y|Z$$

$$P(Y, X_j|Z) \overset{H_0}{=} P(Y|Z) \cdot P(X_j|Z)$$

$$\text{or} \quad P(Y|X_j, Z) \overset{H_0}{=} P(Y|Z)$$

to be continued...

Summary and outlook

the significance test suggested on the random forest website has strange properties:

Summary and outlook

the significance test suggested on the random forest website has strange properties:

- the z-score and power increase in the number of trees and decrease in the sample size

Summary and outlook

the significance test suggested on the random forest website has strange properties:

- the z-score and power increase in the number of trees and decrease in the sample size
- the null hypothesis may not reflect what you wanted

Summary and outlook

the significance test suggested on the random forest website has strange properties:

- the z-score and power increase in the number of trees and decrease in the sample size
- the null hypothesis may not reflect what you wanted

$\Rightarrow$ use conditional permutation scheme

Summary and outlook

the significance test suggested on the random forest website has strange properties:

- the z-score and power increase in the number of trees and decrease in the sample size
- the null hypothesis may not reflect what you wanted

$\Rightarrow$ use conditional permutation scheme

$\Rightarrow$ use distribution over > 1 permutations

Summary and outlook

the significance test suggested on the random forest website has strange properties:

- the z-score and power increase in the number of trees and decrease in the sample size
- the null hypothesis may not reflect what you wanted

  $\Rightarrow$ use conditional permutation scheme

  $\Rightarrow$ use distribution over > 1 permutations

for now: stick to the unscaled importance

Breiman, L. and A. Cutler (2008). Random forests - classification manual (website accessed in 1/2008) http://www.math.usu.edu/∼adele/forests/.

Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics 8:25*.