What We Can Learn from Trees and Forests

Carolin Strobl

Institut für Statistik, LMU München

Today's topics

- variable selection bias
  traditional algorithms for trees and forests artificially prefer variables of certain types
- variable importance
  different types of importance measures and concepts
- outlook: learning about algorithms

Variable selection bias

variable selection in standard classification trees is biased:

numeric variables, variables with many missing values and variables with many categories are preferred

(due to multiple testing and biased entropy estimation $\rightarrow$ Gini index, Strobl et al., 2007)

Variable selection bias

variable selection in standard classification trees is biased:

numeric variables, variables with many missing values and variables with many categories are preferred

(due to multiple testing and biased entropy estimation $\rightarrow$ Gini index, Strobl et al., 2007)

Why is that a problem?

Variable selection bias

the number of categories can be - but is not necessarily - an indicator of the relevance of a predictor variable

- example 1:
  - discretize the continuous variable **age** - would you prefer 2 categories or 10 categories?

Variable selection bias

the number of categories can be - but is not necessarily - an indicator of the relevance of a predictor variable

- example 1:
  - discretize the continuous variable **age** - would you prefer 2 categories or 10 categories?
  - if **age** is informative, more information in retained in 10 categories

Variable selection bias

- example 2:
  - consider **age** in 10 categories vs. **gender** in 2 categories which one is more relevant?

Variable selection bias

- example 2:
  - consider **age** in 10 categories vs. **gender** in 2 categories which one is more relevant?
  - we don't know yet – but it is not necessarily the one with more categories!

Variable selection bias

- example 2:
  - consider **age** in 10 categories vs. **gender** in 2 categories which one is more relevant?
  - we don't know yet – but it is not necessarily the one with more categories!

for trees and forests: need variable selection criteria that are not biased towards certain types of variables

Variable selection bias

biased variable selection criteria for trees

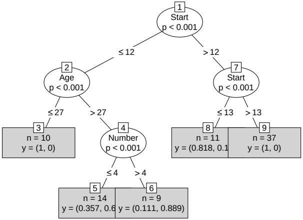
- Gini index as in CART ($\rightsquigarrow$ `rpart`) (Breiman et al., 1984)
- information gain as in C4.5 (Quinlan, 1986)

unbiased variable selection criteria for trees

- ANOVA F-test and χ^2-tests as in QUEST (Loh and Shih, 1997)
- maximally selected statistics (Miller and Siegmund, 1982; Lausen et al., 1994; Shih, 2004; Strobl et al., 2007)
- unbiased entropy estimators (Strobl, 2005)
- conditional inference tests ($\rightarrow$ `ctree`) (Hothorn et al., 2006)

Question

(un)biased variable selection
and variable importance
in classification trees

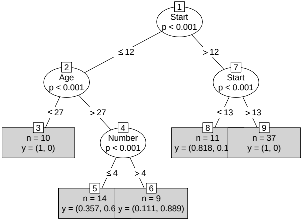


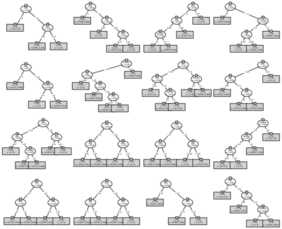
Question

(un)biased variable selection
and variable importance
in classification trees

$\Rightarrow$

(un)biased variable selection
and variable importance
in random forests?





Variable selection and variable importance bias in random forests

- Gini importance (`randomForest`)
  mean Gini gain produced by X_j over all trees

- permutation importance (`randomForest, cforest`)
  mean decrease in classification accuracy after permuting X_j over all trees

Variable selection and variable importance bias in random forests

- Gini importance (`randomForest`)
  mean Gini gain produced by X_j over all trees

  – biased in favor numeric variables and variables with many categories

- permutation importance (`randomForest, cforest`)
  mean decrease in classification accuracy after permuting X_j over all trees

Variable selection and variable importance bias in random forests

- Gini importance (`randomForest`)
  mean Gini gain produced by X_j over all trees

  – biased in favor numeric variables and variables with many categories

- permutation importance (`randomForest, cforest`)
  mean decrease in classification accuracy after permuting X_j over all trees

  + unbiased only if

    1. unbiased variable selection criteria and
    2. subsampling without replacement

  are used, as is default in `cforest` (Strobl et al., 2007)

Variable selection and variable importance bias in random forests

- Gini importance (`randomForest`)
  mean Gini gain produced by X_j over all trees

  – biased in favor numeric variables and variables with many categories

- permutation importance (`randomForest, cforest`)
  mean decrease in classification accuracy after permuting X_j over all trees

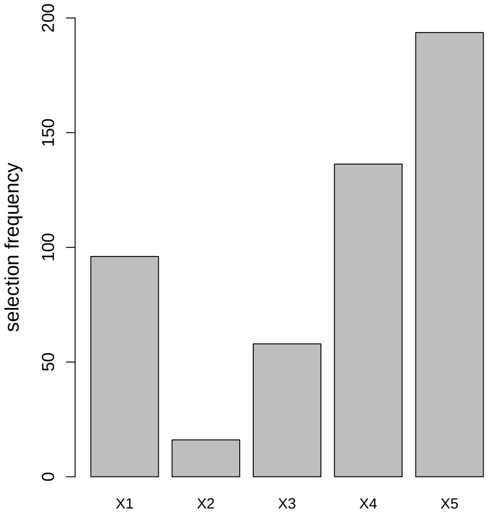
  + unbiased only if
    1. unbiased variable selection criteria and
    2. subsampling without replacement

  are used, as is default in `cforest` (Strobl et al., 2007)

- same for variable selection frequencies

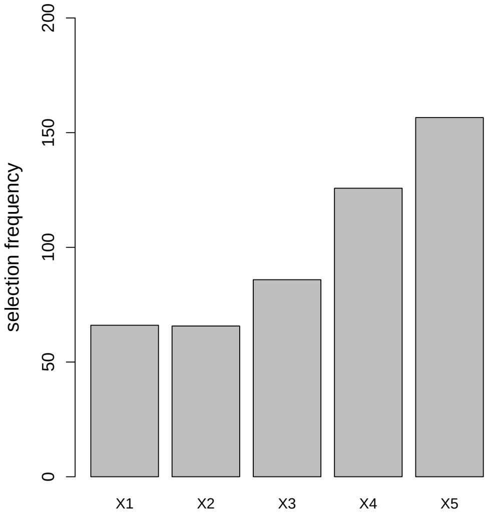
Variable selection frequencies

`randomForest` (biased trees, replace = TRUE)



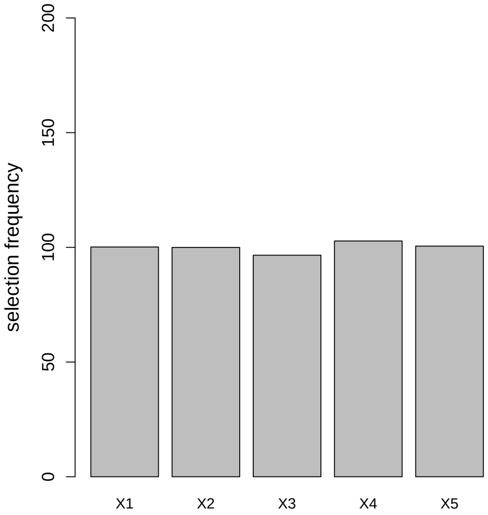
Variable selection frequencies

`cforest` (unbiased trees, replace = TRUE)



Variable selection frequencies

`cforest` (unbiased trees, replace = FALSE)



Variable importance concepts

variable selection in trees and forests is “marginal”

permutation importance is “marginal”

Variable importance concepts

variable selection in trees and forests is “marginal”

permutation importance is “marginal”

Why is that a problem?

Variable importance concepts

example:

in samples of school-children

- shoe size is highly correlated with reading skills
- unless you control for age...

Variable importance concepts

example:

in samples of school-children

- shoe size is highly correlated with reading skills
- unless you control for age...

Variable importance concepts

- marginal correlations
- partial correlations, standardized betas
  conditional effects of X_j given all other variables in the model
- "averaging over orderings"
  - for linear models (`relaimpo`, Grömping, 2006)
    LMG Lindeman, Merenda, and Gold (1980),
    $\approx$ "dominance analysis" Azen and Budescu (2003)

  R^2 decomposition
- random forest permutation importance
  $\approx$ "averaging over trees"

Desirable (?) properties

- *proper decomposition*: scores sum up to model R^2
- *non-negativity*
- *exclusion*: $\beta_j = 0 \Rightarrow \text{score} = 0$
- *inclusion*: $\beta_j \neq 0 \Rightarrow \text{score} \neq 0$

Grömping (2007)

Desirable (?) properties

- *proper decomposition*: scores sum up to model R^2
  LMG
- *non-negativity*
- *exclusion*: $\beta_j = 0 \Rightarrow \text{score} = 0$
- *inclusion*: $\beta_j \neq 0 \Rightarrow \text{score} \neq 0$

Grömping (2007)

Desirable (?) properties

- *proper decomposition*: scores sum up to model R^2
  LMG
- *non-negativity*
  LMG, RF varimp (on average)
- *exclusion*: $\beta_j = 0 \Rightarrow \text{score} = 0$
- *inclusion*: $\beta_j \neq 0 \Rightarrow \text{score} \neq 0$

Grömping (2007)

Desirable (?) properties

- *proper decomposition*: scores sum up to model R^2
  LMG
- *non-negativity*
  LMG, RF varimp (on average)
- *exclusion*: $\beta_j = 0 \Rightarrow \text{score} = 0$
  partial correlations, standardized betas,
  RF varimp?
- *inclusion*: $\beta_j \neq 0 \Rightarrow \text{score} \neq 0$

Grömping (2007)

Desirable (?) properties

- *proper decomposition*: scores sum up to model R^2
  LMG
- *non-negativity*
  LMG, RF varimp (on average)
- *exclusion*: $\beta_j = 0 \Rightarrow \text{score} = 0$
  partial correlations, standardized betas,
  RF varimp?
- *inclusion*: $\beta_j \neq 0 \Rightarrow \text{score} \neq 0$
  all

Grömping (2007)

Simulation study

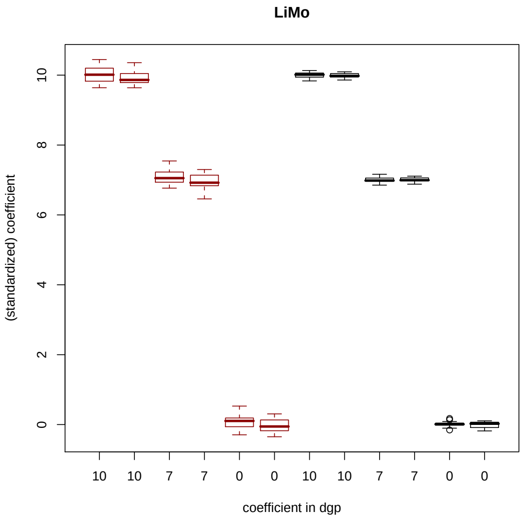
dgp: $y_i = \beta_1 \cdot x_{i,1} + \cdots + \beta_{12} \cdot x_{i,12} + \varepsilon_i,\ \varepsilon_i \overset{i.i.d.}{\sim} N(0,1)$

$X_1, \ldots, X_{12} \sim N(0, \Sigma)$

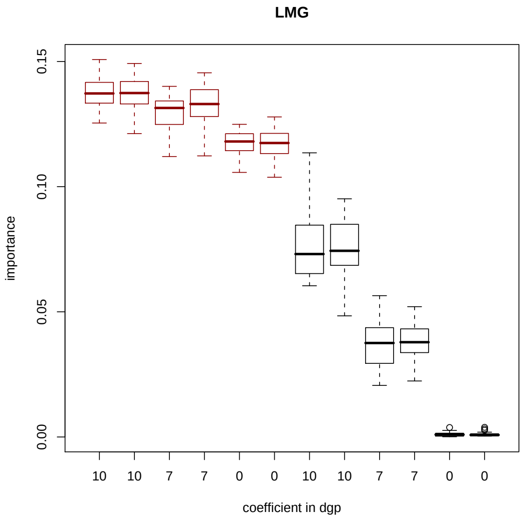
$$\Sigma = \begin{pmatrix} 1 & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & 0 & \cdots & 0 \\ \mathbf{0.9} & 1 & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & 0 & \cdots & 0 \\ \mathbf{0.9} & \mathbf{0.9} & 1 & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & 0 & \cdots & 0 \\ \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & 1 & \mathbf{0.9} & \mathbf{0.9} & 0 & \cdots & 0 \\ \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & 1 & \mathbf{0.9} & 0 & \cdots & 0 \\ \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & \mathbf{0.9} & 1 & 0 & \cdots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & \ddots & \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

X_j	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	$\mathbf{X_5}$	$\mathbf{X_6}$	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
β_j	**10**	**10**	**7**	**7**	**0**	**0**	10	10	7	7	0	0

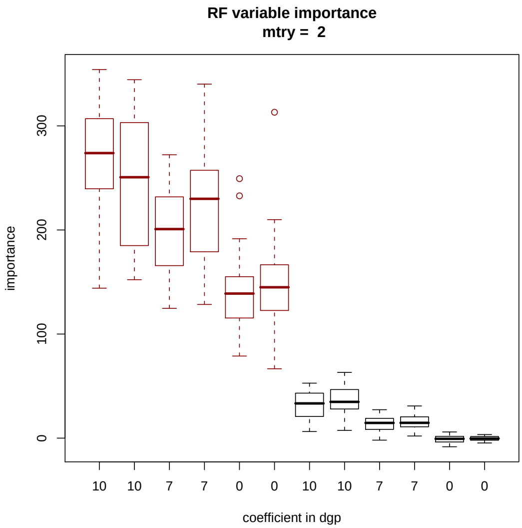
Linear model



LMG


LMG
importance
0.00
0.05
0.10
0.15
10 10 7 7 0 0 10 10 7 7 0 0
coefficient in dgp

RF permutation importance



RF permutation importance

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

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

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

Suggestion: conditional permutation importance

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}	$x_{\pi_{j\|Z=b}(33),j}$	$z_{33} = b$
$\vdots$	$\vdots$	$\vdots$	$\vdots$

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

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

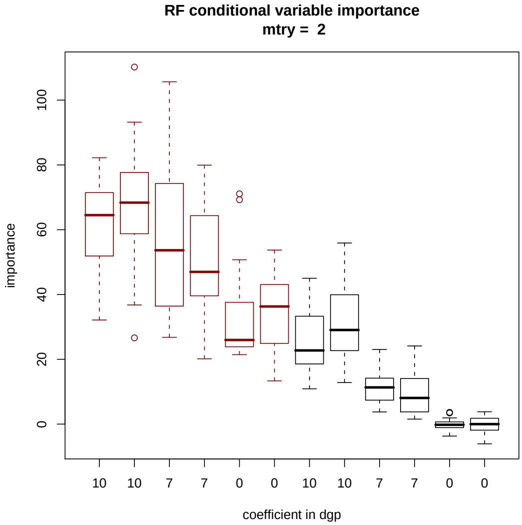
or $$P(Y|X_j, Z) \overset{H_0}{=} P(Y|Z)$$

Example: conditional permutation importance

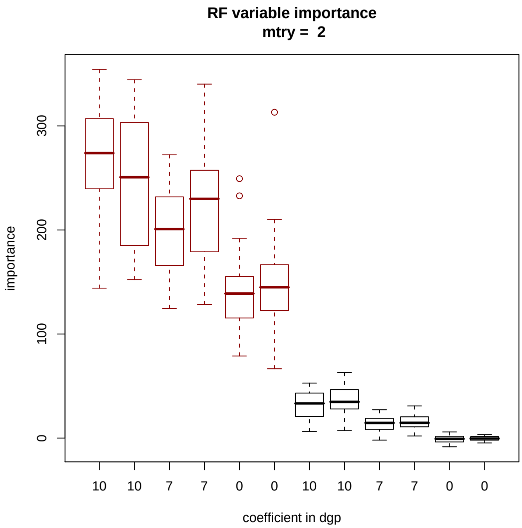
spurious correlation between shoe size and reading skills in school-children

```
> mycf <- cforest(score ~ ., data = readingSkills,
+                  control = cforest_unbiased(mtry = 2))

> varimp(mycf)
nativeSpeaker           age      shoeSize
     12.62926      74.89542      20.01108

> varimp(mycf, conditional = TRUE)
nativeSpeaker           age      shoeSize
    11.808192     46.995336      2.092454
```

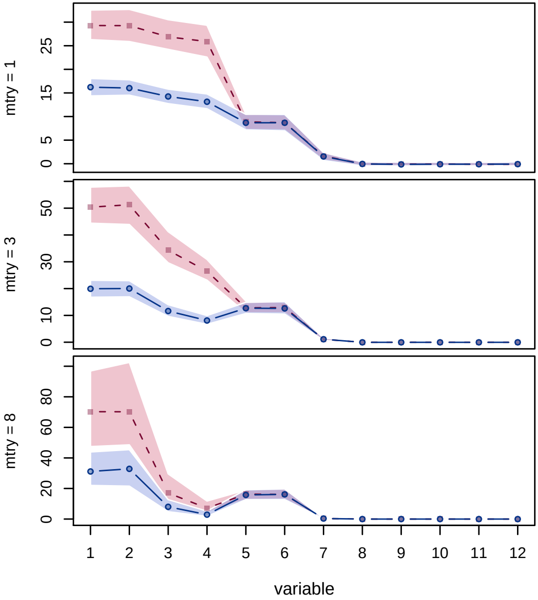
RF conditional permutation importance



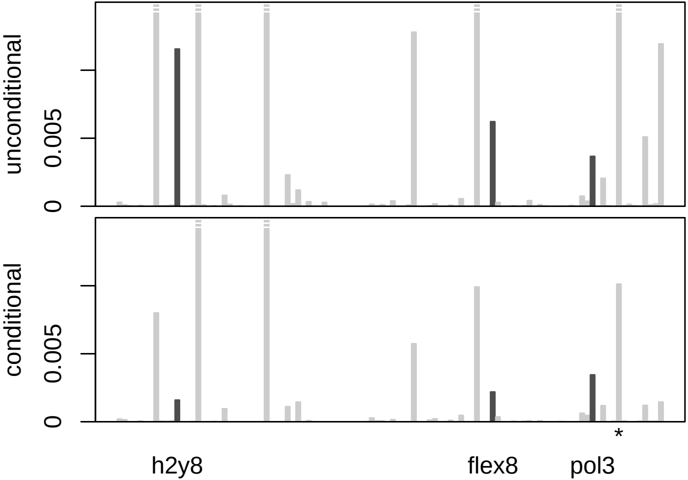
RF unconditional permutation importance



Permutation importance



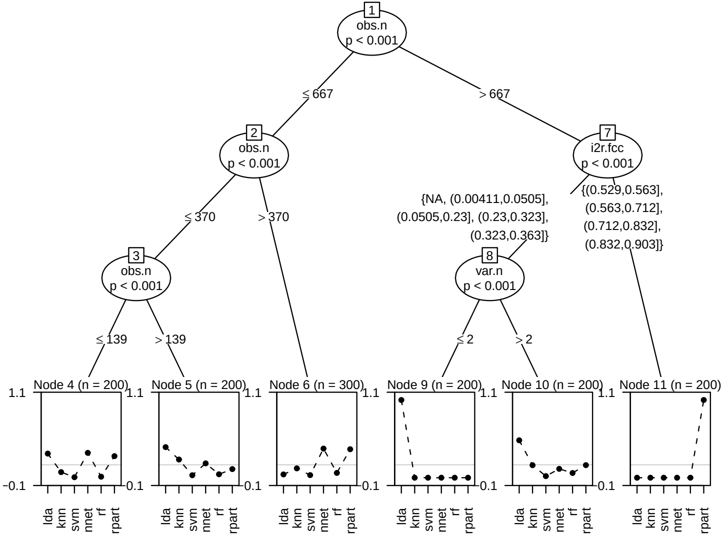
Peptide-binding data



Conclusion

- variable selection bias:
  - affects traditional algorithms for trees and forests
  - use unbiased criteria and subsampling without replacement to avoid bias (as in `cforest`)
- variable importance:
  - conditional permutation importance is computationally expensive and by no means perfect, but more closely resembles partial correlations – if that is what you want
- advantages of random forest variable importance:
  - applicable in high-dimensional settings
  - detect nonlinear and interaction effects

Outlook: use trees to learn about algorithms



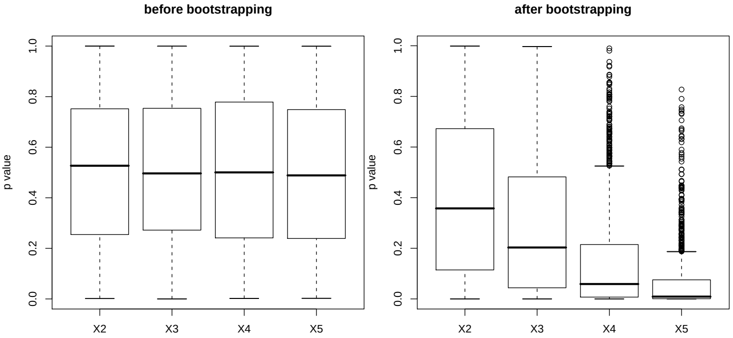
Strobl, C., A.-L. Boulesteix, A. Zeileis, and T. Hothorn (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics 8:25*.

Strobl, C., A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics 9:307*.

Eugster, M., Leisch, F., and Strobl, C. (2010). (Psycho-)Analysis of Benchmark Experiments. A Formal Framework for Investigating the Relationship between Data Sets and Learning Algorithms. *LMU Department of Statistics: Technical Reports, No.78*.

Bootstrap bias

distribution of the p-values of a χ^2-test before and after bootstrapping (1000 iterations with n = 10 000)



Bootstrap bias

- bootstrap sampling with replacement artificially induces an association
- the effect is more pronounced for contingency tables with many df

$\Rightarrow$ in random forests: variables with many categories are again preferred

Bootstrap bias

- for bootstrap testing
  - compute statistic from original sample
  - bootstrap distribution from sample adjusted for the null hypothesis

Bootstrap bias

- for bootstrap testing
  - compute statistic from original sample
  - bootstrap distribution from sample adjusted for the null hypothesis
- here
  - compute statistic from unadjusted bootstrap sample
  - deviation from the null hypothesis increases with df