MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

On Oblique Random Forests

Bjoern Menze, Michael Kelm, Daniel Splitthoff, Ullrich Koethe, Fred Hamprecht

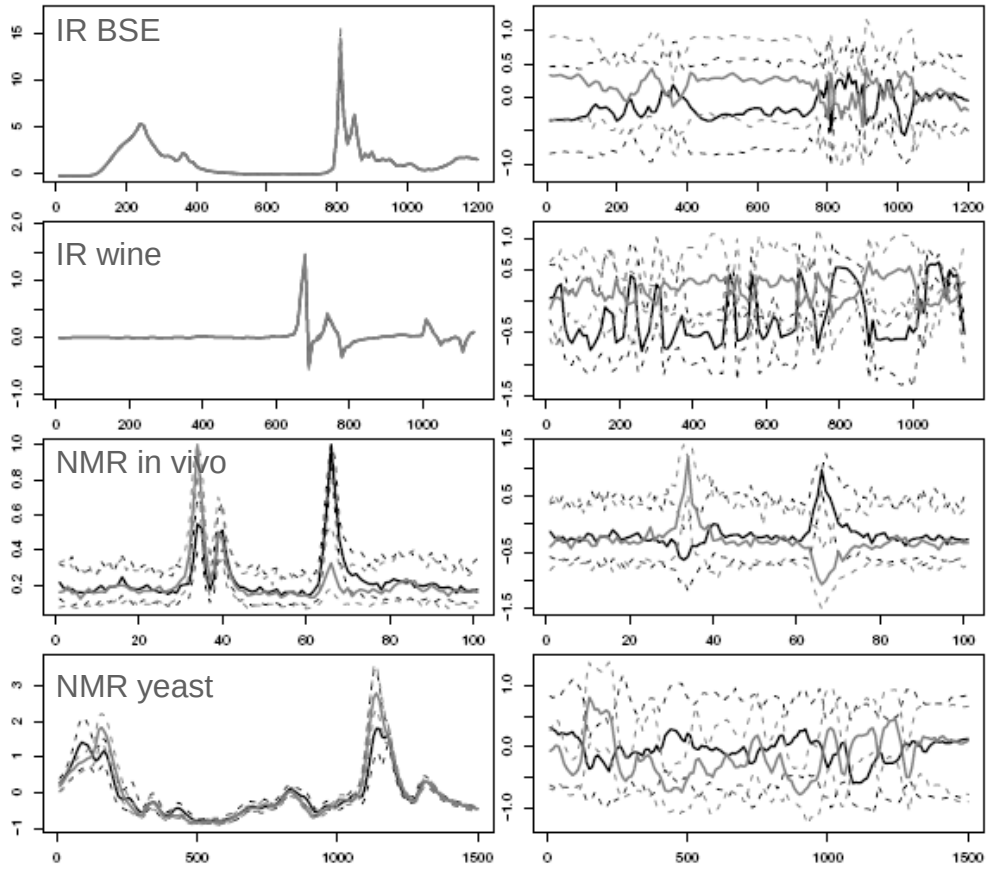
Interdisciplinary Center for Scientific Computing (IWR)
University of Heidelberg, Germany

Computer Science and Artificial Intelligence Laboratory (CSAIL)
Massachusetts Institute of Technology, Cambridge, USA

Overview

- Introduction
- Oblique Random Forest
- Experiment: Performance
- Experiment: Properties
- oRF Tools
- Conclusions

Spectral data



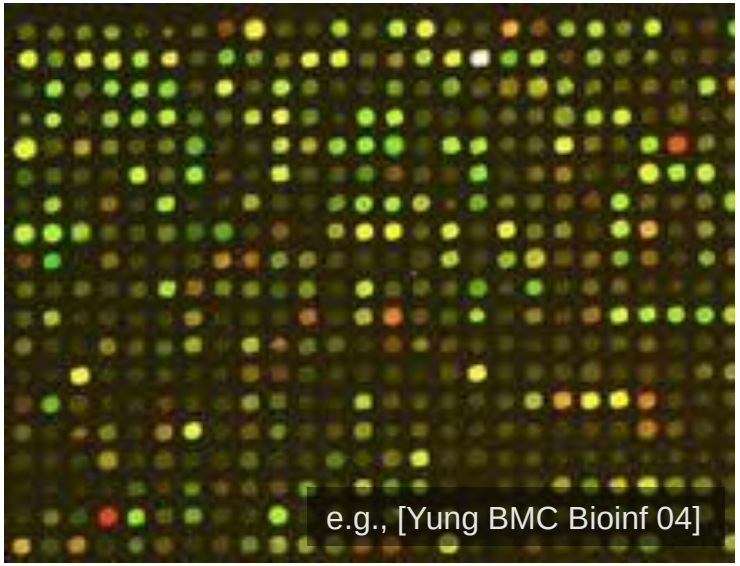
average pattern average pattern w/ normalization

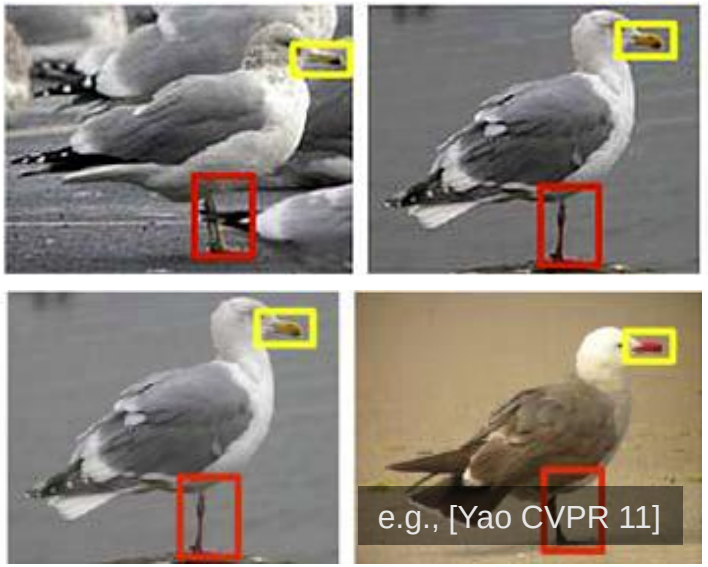
Data properties

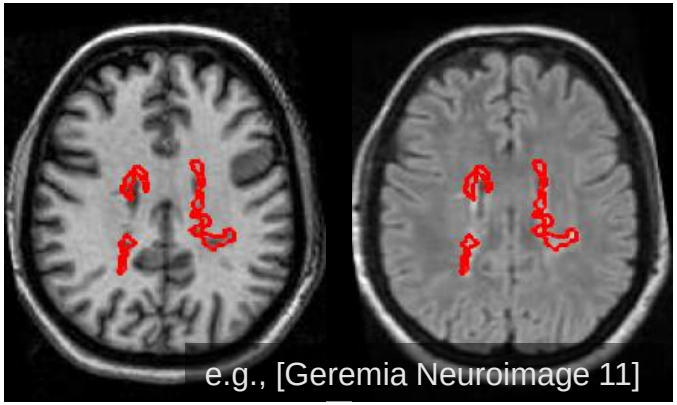
- high-dimensional
- (few samples)
- correlated feature values
- noise → random local offsets

[Menze BMC Bioinformatics 09]

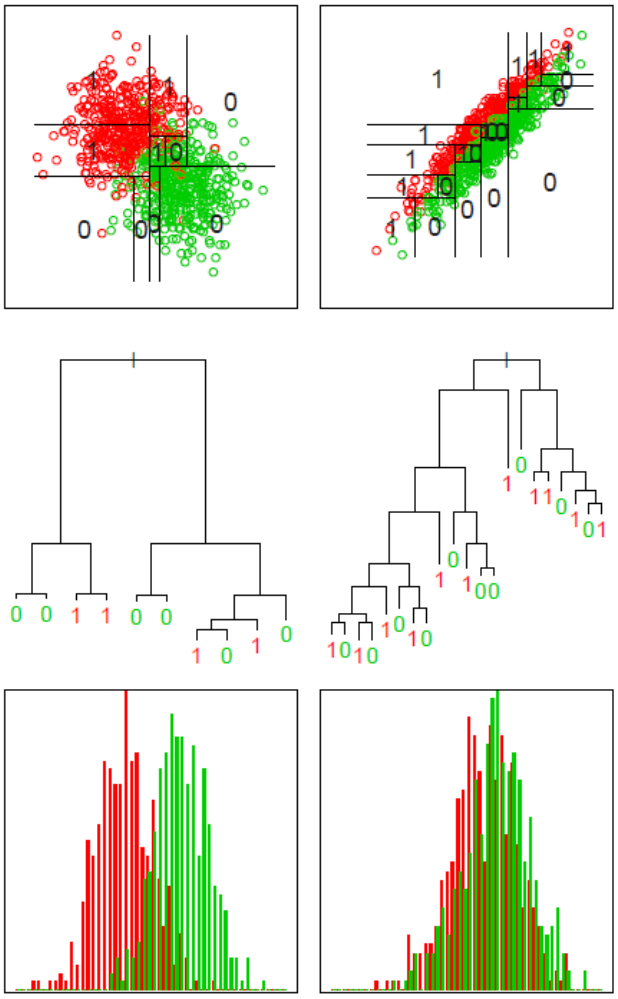
“Spectral data”







Problem: orthogonal splits



Problem:

- Base learner:
  trees w/ orthogonal splits
- Correlated feature values
  → inseparable distributions
  → deeply nested trees

Solution:

- Alternative base learner:
  trees w/ oblique splits

  (Heath 1993; Brodley 1995
  Friedman 1977; Loh 1988)

Prior work

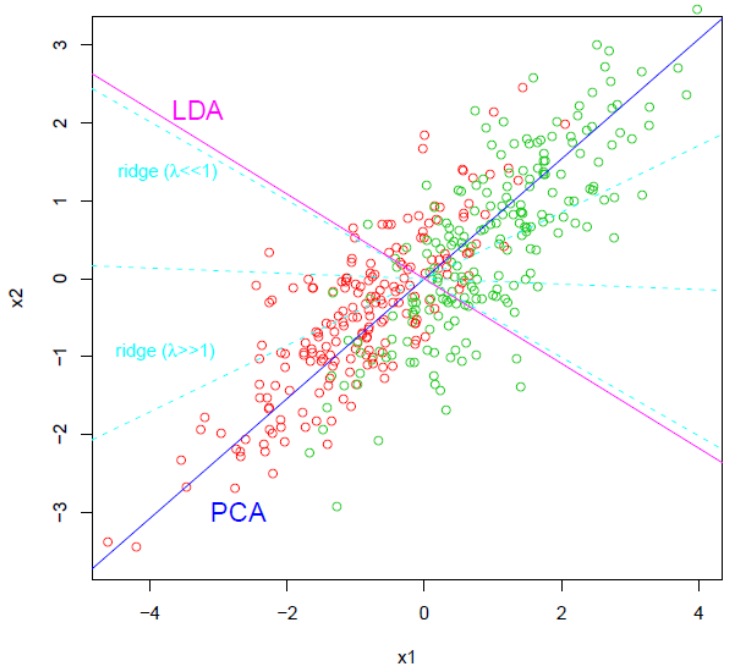
Random decision tree ensembles w/ oblique splits:

- Random forests (Breiman, 2001):
  - random multivariate splits
  - “results never reached so far“ on the data sets tested
- Rotation forests (Rodriguez, 2006):
  - split directions from principal components
  - “improving results significantly in selected classification tasks”
- MML Forest (Tan, 2006):
  - global optimization of all split directions in the tree (under minimum-description-length criterion)
  - “favourable results compared to other ensemble learning algorithms”

Overview

- Introduction
- Oblique Random Forests
- Experiment: Performance
- Experiment: Properties
- oRF Tools
- Conclusions

Oblique model trees



Split directions under different node models

- **Recursive binary splits** :

$$f_m(\mathbf{x}) : \beta_m^T \mathbf{x} > c_m$$

with coefficients β_m and threshold c_m

- **Coefficients** β_m : ridge regression

$$\beta_{ridge}(\lambda) \sim \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{2} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{P} |\beta_j|^2$$

$$\beta_{ridge}(\lambda') \sim \underset{||\beta=1||}{\operatorname{argmax}} corr^2(\beta X, Y) * \frac{var(\beta X)}{var(\beta X) + \lambda'}$$

- **Threshold** c_m : e.g., max. Gini decrease

Oblique random forest

Oblique random forest.

---

1. Choose the number of trees to grow (n_{tree}).
2. Choose the number of variables used to split each node (m_{try}).
3. Grow n_{tree} as follows:
   (a) Draw a bootstrap sample of size n (with replacement) and grow a tree from this bootstrap sample.
   (b) When growing a tree, select m_{try} variables at random at each node. Find the best split direction β in the subspace spanned by these variables according to the predefined node model. (If your node model has hyper-parameters λ, tune them on the out-of-bag samples available at that node.) The model fitted values are used to identify the threshold c for the binary split.
   (c) Grow the tree to full depth. (Do not prune.)
4. In prediction classify a new sample according to the majority of the predictions collected from all trees in the forest.

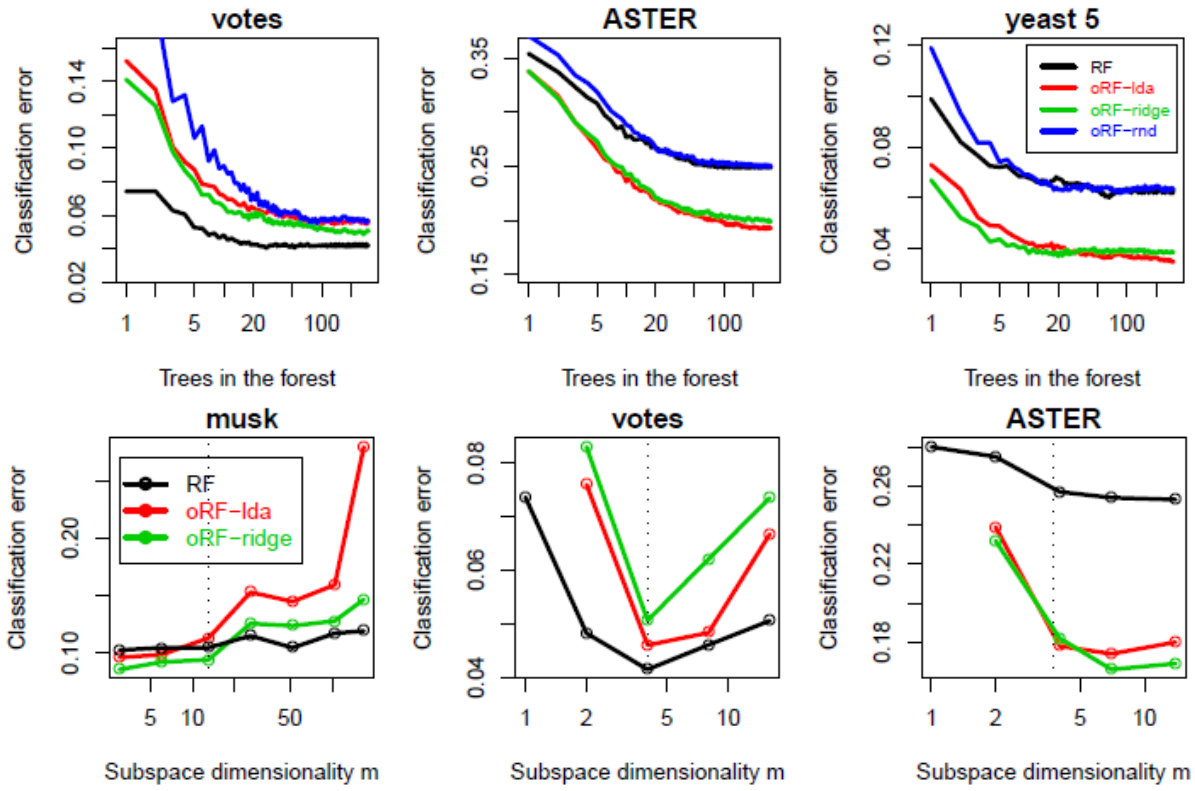
Scaling behavior

- Computational complexity scales linearly with the number of trees in the ensemble.
- P: subspace dimensionality; N: (sub-) sampled observations

	Training Node	**Training Splits**	**Training Tree**	**Prediction Tree**
RF Tree	P N log N (feature ranking)	log N (N levels)	**P N $(\log N)^2$**	**log N**
oRF Tree	P^2 N [P N t + t^3; t = min (P,N)] (regression)	log N (N levels)	**P^2 N log N**	**P log N**

→ for small P or high N (i.e., P < log N), oRF is more efficient to train than RF

Parameterization



RF
oRF w/ LDA
oRF w/ ridge
oRF w/ random

(10 fold x-val)

- Default : ensemble size n_{tree} = 300 trees
- Default : subspace dimensionality m_{try}= sqrt(P)

Overview

- Introduction
- Oblique Random Forest
- Experiment: Performance
- Experiment: Properties
- oRF Tools
- Conclusions

Comparison

Data

- 40 binary classification tasks, (balanced classes)
- 10 factorial, 10 numerical, 20 spectral data sets

Classifiers

- oRF-rnd, oRF-lda, oRF-ridge → *R package* ***obliqueRF***
  w/ default parameters
- SVM (RBF), knn, CART, RF, RF-rnd, adaboost

  w/ parameters optimized in inner x-val

Measures

- AUC ROC, Accuracy
- 10* 10 fold x-val

Performance

		svm	knn	CART	adaboost	RF	RF-rnd	oRF-rnd	oRF-lda	oRF-ridge	svm	knn	CART	adaboost	RF	RF-rnd	oRF-rnd	oRF-lda	oRF-ridge
1	chess	2.8	5.4	2.4	**0.6**	1.4	52.2	23.5	4.1	2.7	0.3	1.2	0.9	**0**	0.1	33.3	0.9	0.4	0.3
2	credit	15.3	32	14.8	**12.9**	**12.7**	17.9	15.4	**13**	44.5	10.3	27.3	15.4	**6.7**	**6.8**	7.5	8.3	**7**	53.3
3	heart	23	35.5	21.3	**18.3**	**17.8**	40.6	20	**18**	**17.9**	15.7	30.9	20.5	11.4	10.4	13.5	12.7	**9.9**	**9.9**
4	hepatitis	20.9	21.6	**20.3**	**24.4**	**17.8**	**18.7**	**19.1**	**19.3**	20.6	47	42.2	28.7	18.9	**14.6**	19.5	16.2	15.7	50
5	monks3	3.6	**1.6**	**1.1**	**1.2**	2.4	43.3	5.7	3.5	**1.3**	1.4	**0.9**	1.7	**0.8**	1.1	5.8	**0.8**	1.1	1
6	promotergene	17.1	20.3	30.4	**9.6**	**11**	50.1	20.8	16.9	18.5	8.8	12	31.3	**3.9**	**4**	32	12.8	7.3	7.4
7	tic-tac-toe	12	16.1	26.5	**0**	4.7	65.3	12.7	15.2	14.8	4	32	31.1	**0**	0.3	73.3	4.9	3.5	2.8
8	titanic	**21.6**	23.3	23	**21.3**	22.7	30.2	26.8	22	24.7	20.1	20.1	25.5	**16.1**	16.8	22.1	20.3	17.6	18.6
9	votes	9.3	12.5	**4.4**	**4.5**	**4.2**	7.7	5.7	5.6	5.1	4.2	8.7	6.1	1.4	**0.9**	1.8	1.4	1.4	1.3
10	cancer	4.3	**3.4**	6.2	3.7	**3.3**	**3**	**2.8**	**3**	**3**	2.3	1.3	4.8	1	0.9	0.9	**0.7**	0.9	0.9
11	ionosphere	**5.1**	14	10.5	**6.4**	**6.5**	7.5	**5.6**	**5.4**	**5.6**	2	16.9	12.3	3.3	2.1	2.1	**1.7**	**1.7**	**1.6**
12	sonar	17.5	18.2	27.2	**13.1**	**16.1**	16.3	**14.4**	18	18.2	7.6	17.3	25.7	**6.1**	6.8	**6.4**	**5.8**	6.5	**6.2**
13	musk	10.4	14.6	19.6	**8.8**	10.6	17.9	10.2	10.6	10.4	3.8	12.9	15.9	**2.9**	4.4	5.3	3.7	5	4.5
14	liver	30.9	32	34.1	**26.8**	**26.8**	**27.8**	27.8	**25.8**	**26.2**	26.6	31.3	38.2	22.3	23.3	23.5	23.4	**21.4**	**21.3**
15	diabetes	30.7	31.7	34.2	**26.6**	**26.7**	**27**	**27.7**	**26.1**	**26.2**	26.8	30	37.7	22.3	23.3	23.3	23.1	**21.2**	**21.5**
16	ringnorm	**12.4**	17	19.8	13.2	17.5	17.9	18	18	18	**2.1**	25.5	40.9	7.2	24.7	14.2	17.9	36.3	35.4
17	spirals	**5.2**	**5.1**	10.3	6.1	6	6.1	**5.5**	**5.3**	**5.2**	**1.1**	1.5	6.9	2.2	2.1	2.1	1.3	1.4	1.3
18	threenorm	**13.3**	14.9	33.8	15.3	16.4	15.1	**14**	14.6	14.7	**6**	6.6	30.6	7.7	7.8	7.4	6.7	7.2	7
19	twonorm	**2.2**	**1.9**	21.6	3.4	3.5	2.8	**2.2**	**1.8**	**1.7**	0.2	0.2	18.5	0.4	0.5	0.3	0.2	**0.2**	**0.2**
20	circles	**2.1**	2.5	5.1	**2.4**	2.8	3.3	**1.7**	**1.9**	**1.8**	0.1	1.8	3.1	0.3	0.4	0.4	**0**	**0**	**0**
21	digits 2-4	1.3	**0.4**	5	2.4	1.7	50	1.5	2.9	2.7	**0.1**	0.3	3.4	0.3	0.1	5.9	0.1	0.5	0.6
22	digits 3-8	**2.3**	3.6	6	**3.1**	**3.2**	50	4.3	4.9	5.2	**0.3**	2.3	4.4	0.6	0.5	30.7	0.8	1.2	1.3
23	digits even-odd	**7.4**	**7.2**	15.8	9.1	**7.8**	50	9.8	16.2	16.6	**2.4**	6.2	14.6	3.2	**2.3**	47.2	3.3	9	9.4
24	RS SRTM	2.5	3.2	3	**0.8**	**0.7**	1.7	1.7	**0.7**	1	0.3	2.7	2.7	0.1	**0**	0.1	0.1	**0**	**0**
25	RS ASTER	25.5	29.8	32.7	22.9	24.9	27.5	25.1	**19.3**	**19.9**	18.7	25.4	30.1	17.4	19.6	20.8	18.6	**13.2**	14
26	MRS quality	**6.2**	6.7	18.1	6.7	7.7	9.3	8.4	**6.2**	**6.2**	2.1	2.5	15.5	2.4	2.1	2.7	2.4	1.9	**1.9**
27	MRS tumor 1	**11.2**	12	17.5	**10.6**	**10.4**	12.2	**10.7**	**10.6**	**11**	4.9	8.4	16.8	6.2	**4.6**	5.9	4.9	**4.6**	**4.7**
28	MRS tumor 1	**19.1**	**19.7**	22.3	**18.9**	**19**	**19.1**	**19.1**	**19.7**	**20.1**	28	29.5	35.4	**26.8**	**24.7**	**26**	25.9	26.9	**25.5**
29	IR BSE 1	22.5	23.1	25.2	22.1	20.5	23	23	**13.1**	**13.4**	22	33.6	39.3	27.4	21.3	26.8	24.5	**9.2**	9.9
30	IR BSE 2	27.9	42.9	24.1	25.9	25.7	29.6	34.5	**15**	**15.5**	20.5	41.5	30	20.7	18.5	25.1	27.3	**6.9**	**7.1**
31	IR BSE 3	27.1	40.5	25.2	33.5	24.6	30.8	35.4	**14.9**	**12.6**	20.3	39	31.5	18.2	18.5	26.1	29	**5**	**5.2**
32	MRS yeast 1	**4.3**	9	14.5	7.8	7.3	8.9	7.9	**4.1**	**4.3**	**1.2**	6.4	15.5	3.4	2.9	3.6	2.9	**1.2**	**1.2**
33	MRS yeast 2	**2.4**	**3.2**	9	8.3	3.9	4.4	**3.5**	**3**	**2.8**	**2.2**	4.3	11.1	4.5	3.6	2.8	2.9	2.7	2.8
34	MRS yeast 3	**3**	4.9	8.7	7.2	5	5.2	4.4	**3.2**	**3.2**	3.9	4.8	13.6	9.5	4.8	5	4.8	**3.2**	**3.2**
35	MRS yeast 4	9.7	11	15.7	15.8	12	14.1	13.3	**6.5**	**5.9**	**4.2**	14.2	26.9	6.1	8.6	8	7.1	4.5	**4.2**
36	MRS yeast 5	5	7.1	8	7.4	6.4	6.4	6.4	**3.5**	**3.9**	**3.2**	8.6	16.2	9.9	4.3	5.6	5.5	**3.4**	**3.1**
37	IR wine origin 1	**27.2**	40.7	**26**	**22.1**	**21.7**	**26.4**	29.6	**21.4**	**21.6**	23.5	40.6	28	16.9	**14.5**	19.6	23.8	**13.6**	**13.2**
38	IR wine origin 2	**25.5**	40.6	30.3	**21.8**	**21.1**	**25.4**	32.1	**25.5**	**22.6**	23.2	41.7	31.9	27.2	**15**	19.2	27.5	15.6	**13.9**
39	IR wine grape 1	17.1	40.3	14.7	18	**11.1**	21.9	25.1	**8.4**	**4.6**	6	38.4	18.8	7.6	3	10.2	16.7	**0**	**0**
40	IR wine grape 2	18	38.2	**15.3**	**12.5**	**10.3**	22.1	29.5	**11.6**	**11.1**	6.2	35.2	18.3	5.5	2.7	10.1	21.2	**0**	**0**

avg. classification error

avg. 1- AUC ROC

- **Underlined:** best performance on given data set (on average)
- **Bold type:** no significant difference to best method in pair-wise comparison of hold out predictions (Cox-Wilcoxon test, .05 significance level)

→ **"best or comparable to the best"**

Ranking

Rank		1	2	3	4	5
Factorial data	AUC ROC Accuracy	adaboost (6) adaboost (9)	RF(4) RF (6)	oRF-lda (2) oRF-lda (4)	oRF-rnd (2) oRF-ridge (3)	knn (1) CART (3)
Numerical data	AUC ROC Accuracy	oRF-ridge (6) oRF-rnd (7)	oRF-lda (5) oRF-ridge (6)	svm (3) oRF-lda (6)	oRF-rnd (3) adaboost (6)	adaboost (2) svm (6)
Spectral data	AUC ROC Accuracy	oRF-ridge (14) oRF-ridge (17)	oRF-lda (12) oRF-lda (17)	svm (7) svm (10)	RF (6) RF (9)	adaboost (1) adaboost (7)

Grouping results on similar types of data: Number of times a method performs "best or comparable to the best" for the given data type

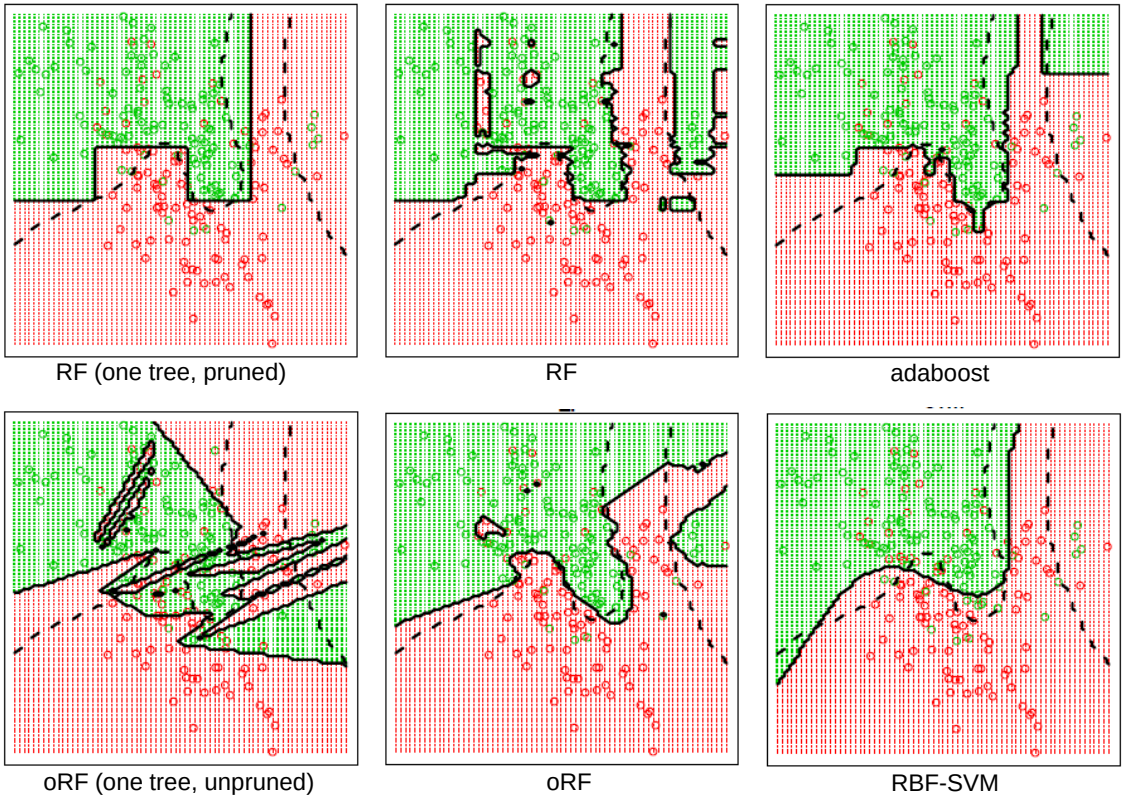
→ **Factorial data:** RF and adaboost best

→ **Nominal, spectral data:** oRF best (lda, ridge)

Overview

- Introduction
- Oblique Random Forest
- Experiment: Performance
- Experiment: Properties
- oRF Tools
- Conclusions

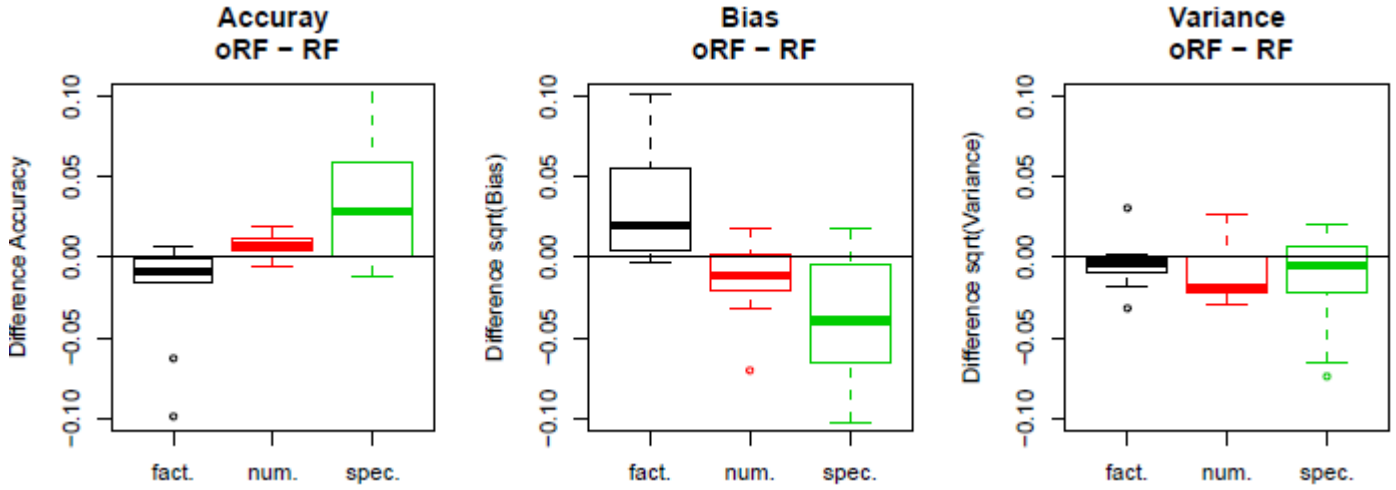
Topology of the decision boundary



Gaussian mixtures data (Hastie et al.)

→ More flexible decision boundary, less structural bias through oblique trees

Bias variance analysis

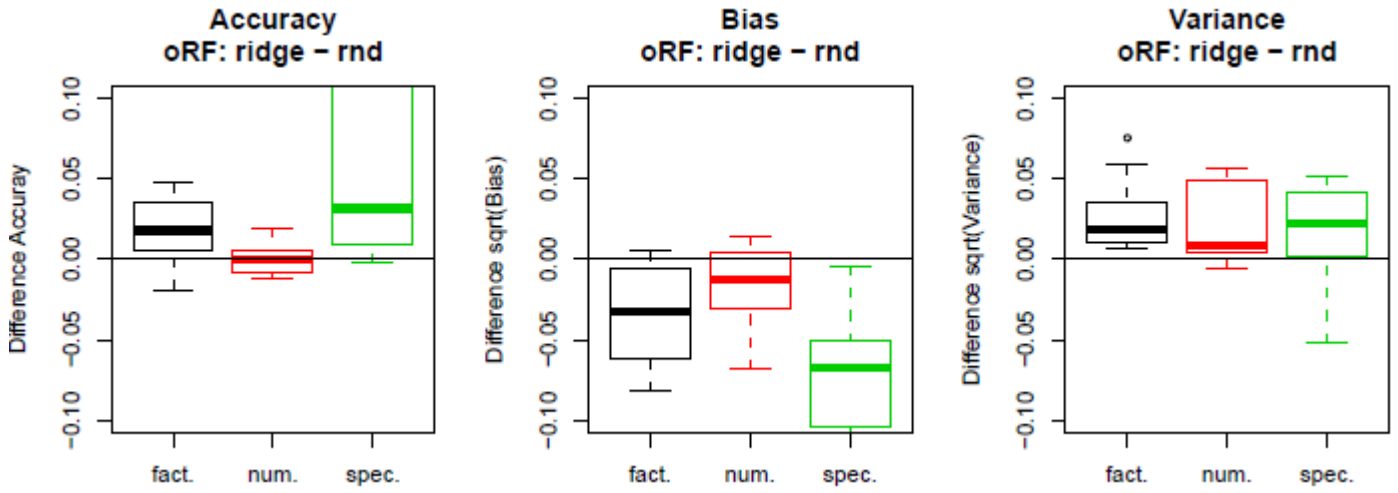


Advantage over RF w/ univariate splits

- RF higher accuracy on factorial data
  oRF higher accuracy on spectral (and num.) data
- oRF and RF with similar variance

→ Performance differences through differences in bias

Bias variance analysis

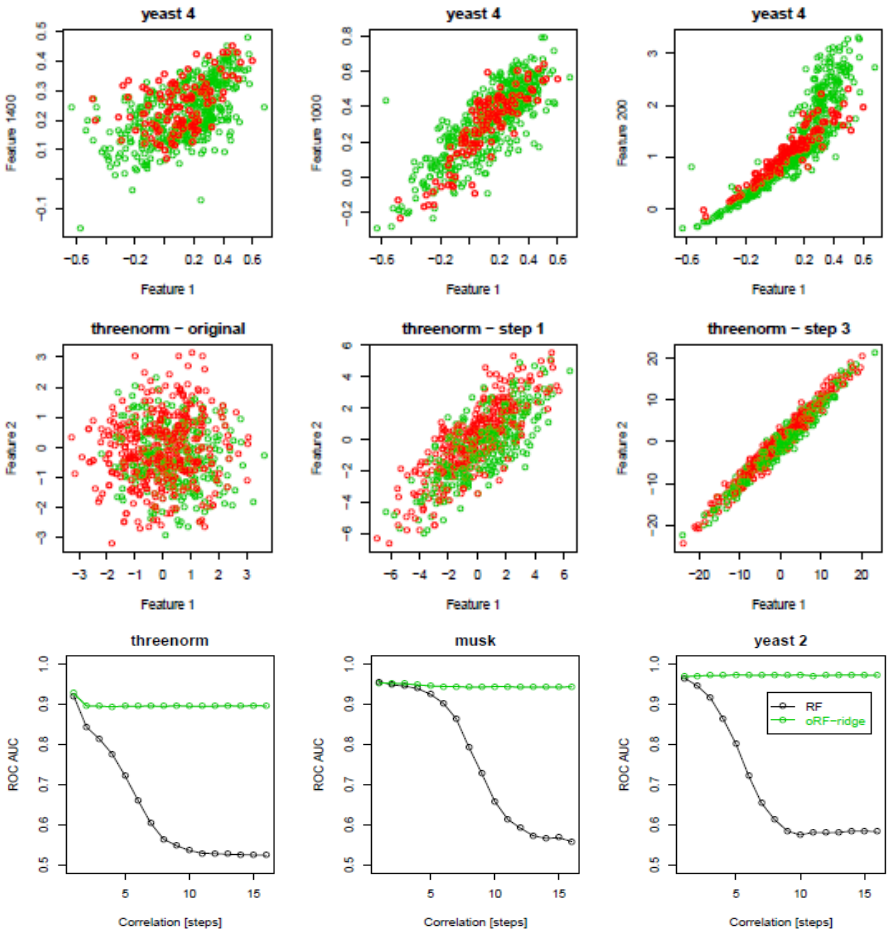


Advantage over oRF w/ multivariate *random* splits

- oRF-ridge higher accuracy
- oRF-rnd lower variance, but largely higher bias

→ Performance differences through differences in bias

Artificially correlated data



Advantage on spectral data

- Random offsets to feature values
- Increases correlation between features

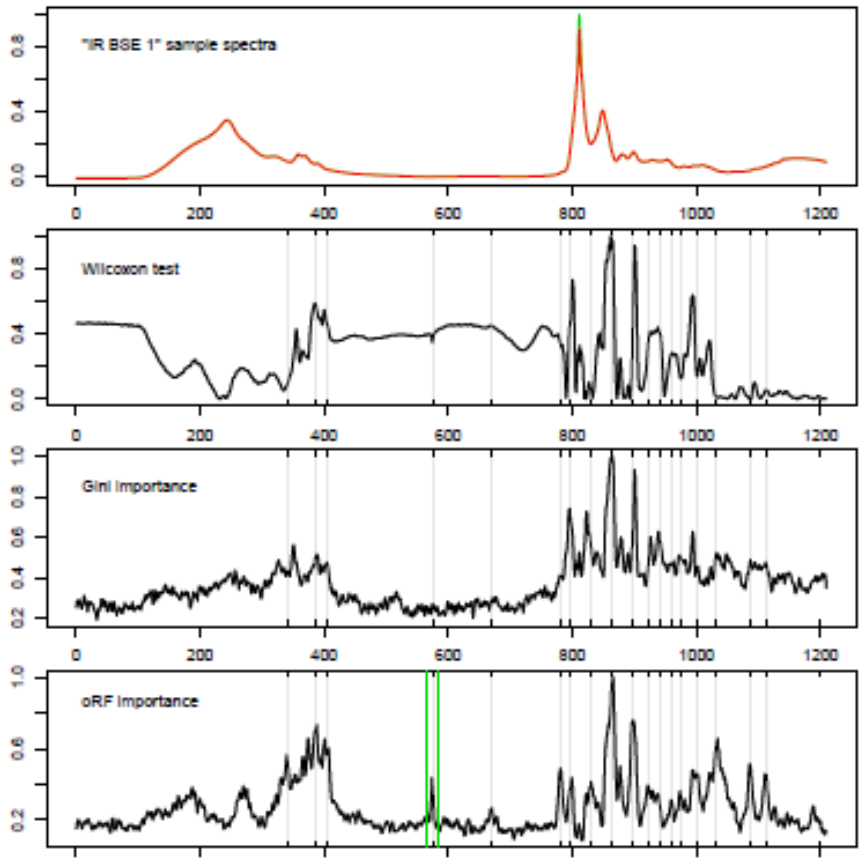
→ RF classification is random for large correlation

→ oRF performance remains (nearly) unchanged

Overview

- Introduction
- Oblique Random Forest
- Experiment: Performance
- Experiment: Properties
- oRF Tools
- Conclusions

Feature importance



Spectral data, feature importance measures

RF feature importance

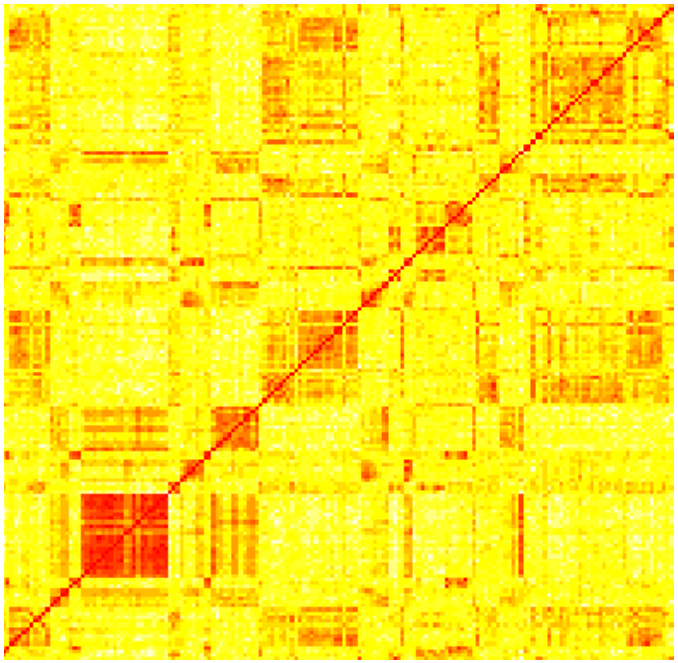
- Statistic of chosen variables ("Gini importance")
- Permutation importance (testing on out-of-bag)

oRF feature importance

- Statistic of variables deemed significant in ANOVA tests at each split

→ oRF importance more detailed; evaluates features "irrelevant" to RF

Sample proximity



Sample proximity

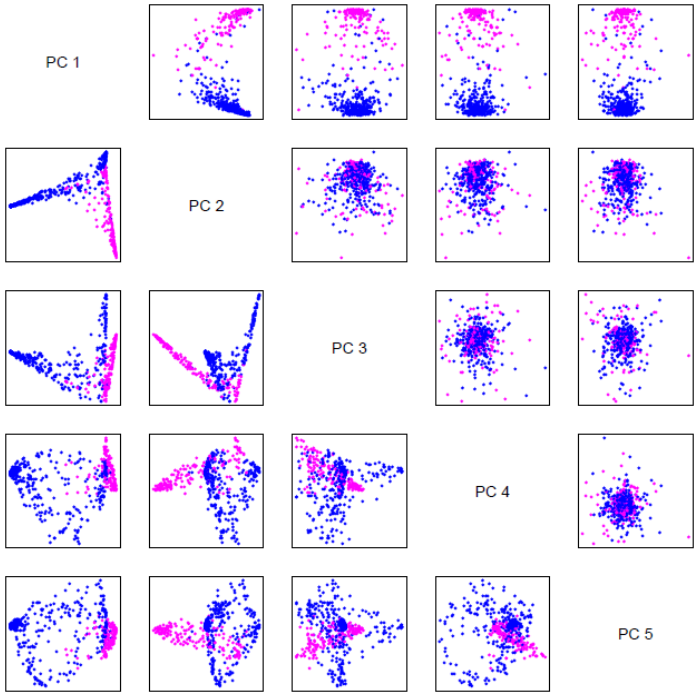
Sample proximity

- Co-occurrence table: counts of how often two out-of-bag samples lie in the same terminal node

→ oRF proximity is void of structures resulting from complex, deeply nested trees

→ oRF proximity explains variation (in MDS embedding) with fewer eigenspaces

Sample proximity



Top five eigen-spaces for MDS embedding of proximity matrix

oRF: upper half; RF lower half

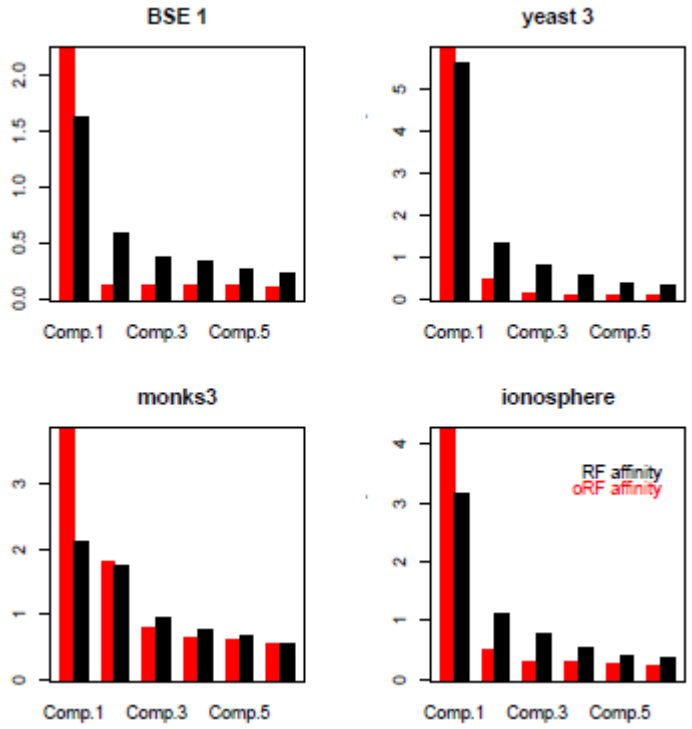
Sample proximity

- Co-occurrence table: counts of how often two out-of-bag samples lie in the same terminal node

→ oRF proximity is void of structures resulting from complex, deeply nested trees

→ oRF proximity explains variation (in MDS embedding) with fewer eigenspaces

Sample proximity



Variance explained in first six eigen-spaces of MDS embedding

oRF: red; RF: black

Sample proximity

- Co-occurrence table: counts of how often two out-of-bag samples lie in the same terminal node

→ oRF proximity is void of structures resulting from complex, deeply nested trees

→ oRF proximity explains variation (in MDS embedding) with fewer eigenspaces

Overview

- Introduction
- Oblique Random Forest
- Experiment: Performance
- Experiment: Properties
- oRF Tools
- Conclusions

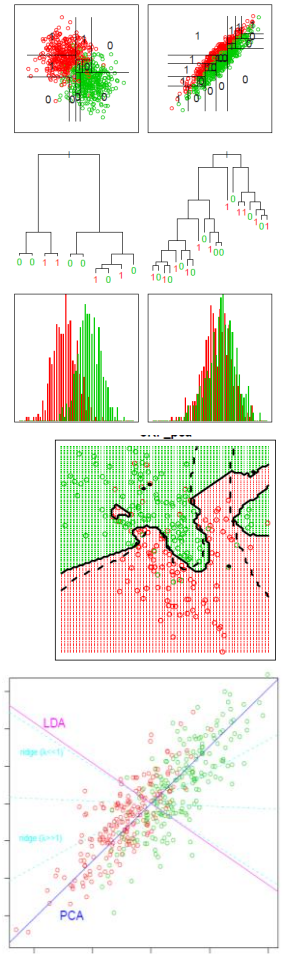
Conclusions

Proposed: oRF with recursive linear model splits

- Efficient scaling behavior through model-splits
- oRF outperforms RF (except on factorial data):

  RF < oRF-rnd < oRF-lda < oRF-ridge
- Advantages oRF: topology; unbiased task-optimal splits; simpler feature importance / proximity measure

→ **More complex node model yields simpler trees and simpler, but better forests**



R package ***obliqueRF*** available at *cran.r-project.org*