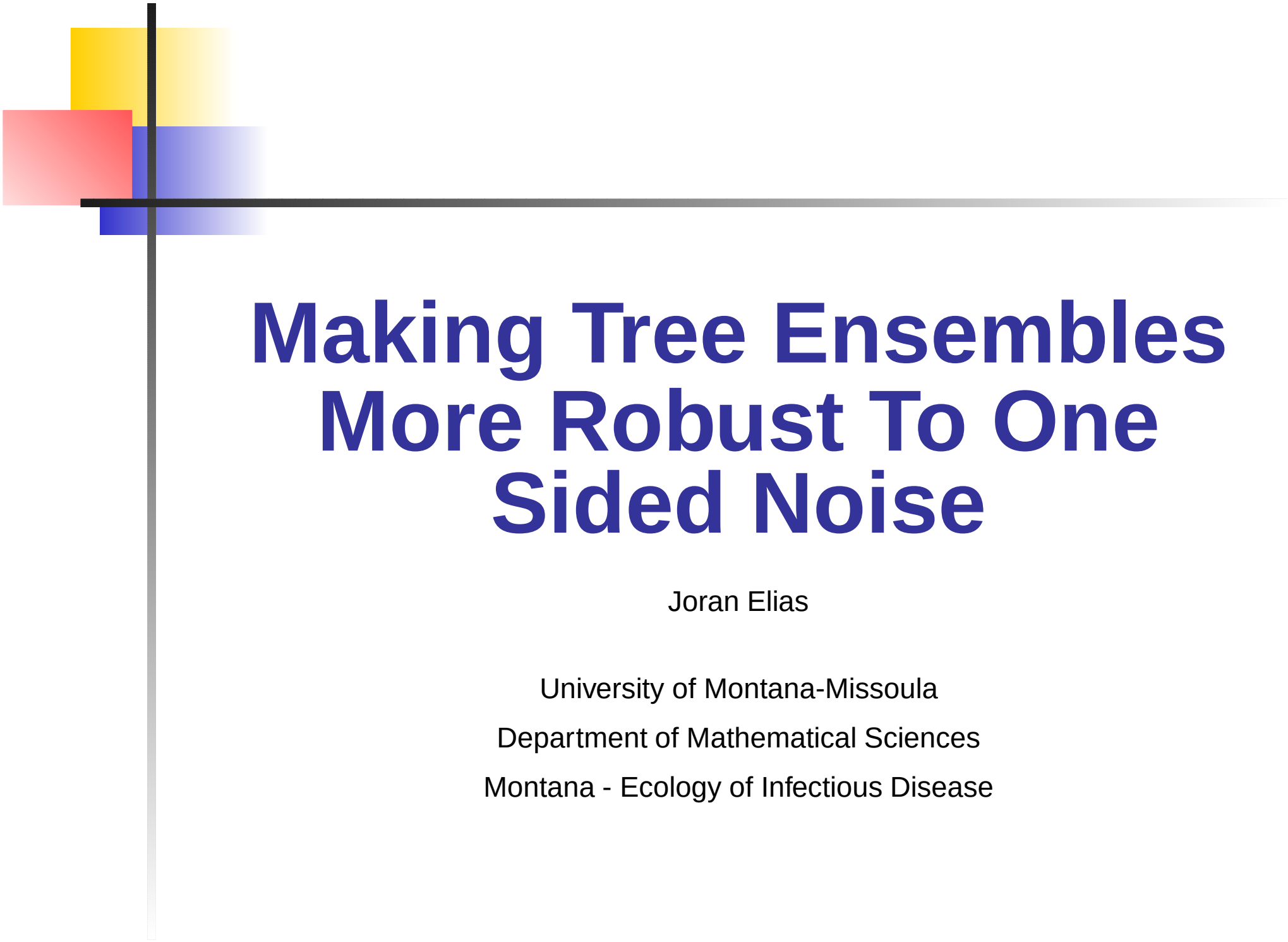


Making Tree Ensembles More Robust To One Sided Noise

Joran Elias

University of Montana-Missoula

Department of Mathematical Sciences

Montana - Ecology of Infectious Disease

Outline

- Introduction
- One Sided Noise
- Sampling Strategy
- Preliminary Simulations
- Conclusion

Tree Ensembles

- Classification trees (*i.e.* CART, C4.5)

  Fits a piecewise constant function by recursively partitioning the feature space

- Tree Ensembles (*i.e.* bagging, Random Forests, etc.)

  Aggregates the predictions from multiple trees

  Some form of randomization is employed when building each tree (*i.e.* bootstrap resampling)

  Only consider two-class classification problems: response variable $y \in \{0, 1\}$ and predictor variables $\mathbf{x}$ (continuous or categorical).

One Sided Noise

- Sometimes we can only observe cases from one class (class 1) and unlabeled cases.

One Sided Noise

- Sometimes we can only observe cases from one class (class 1) and unlabeled cases.
- Naive approach: simply assign all the unlabeled cases to their own class (class 0)

One Sided Noise

- Sometimes we can only observe cases from one class (class 1) and unlabeled cases.
- Naive approach: simply assign all the unlabeled cases to their own class (class 0)
- Observations from one class (class 1) are observed with certainty, observations from the other (class 0) have an unknown level of noise.

One Sided Noise

- Sometimes we can only observe cases from one class (class 1) and unlabeled cases.
- Naive approach: simply assign all the unlabeled cases to their own class (class 0)
- Observations from one class (class 1) are observed with certainty, observations from the other (class 0) have an unknown level of noise.
- Examples include ecological niche modeling and document classification

Sampling Strategy

- Suppose our training sample contains $n = n_1 + n_0$ observations where n_1 are labeled as belonging to class 1 and n_0 are labeled as belonging to class 0.

Sampling Strategy

- Suppose our training sample contains $n = n_1 + n_0$ observations where n_1 are labeled as belonging to class 1 and n_0 are labeled as belonging to class 0.
- Suppose that r of the n_0 class 0 observations are mislabeled (they actually belong to class 1);

Sampling Strategy

- Suppose our training sample contains $n = n_1 + n_0$ observations where n_1 are labeled as belonging to class 1 and n_0 are labeled as belonging to class 0.
- Suppose that r of the n_0 class 0 observations are mislabeled (they actually belong to class 1);
- Random Forest: the expected number of mislabeled observations in each bootstrap sample is r, leading to a constant noise rate of $p_{noise} = r/n$ for each tree in the forest.

Sampling Strategy

- Suppose our training sample contains $n = n_1 + n_0$ observations where n_1 are labeled as belonging to class 1 and n_0 are labeled as belonging to class 0.
- Suppose that r of the n_0 class 0 observations are mislabeled (they actually belong to class 1);
- Random Forest: the expected number of mislabeled observations in each bootstrap sample is r, leading to a constant noise rate of $p_{noise} = r/n$ for each tree in the forest.
- If we can reduce p_{noise} we may see performance improvements...

Sampling Strategy

- Suppose our training sample contains $n = n_1 + n_0$ observations where n_1 are labeled as belonging to class 1 and n_0 are labeled as belonging to class 0.
- Suppose that r of the n_0 class 0 observations are mislabeled (they actually belong to class 1);
- Random Forest: the expected number of mislabeled observations in each bootstrap sample is r, leading to a constant noise rate of $p_{noise} = r/n$ for each tree in the forest.
- If we can reduce p_{noise} we may see performance improvements...
- Select (without replacement) $s < \min\{n_1, n_0\}$ observations from each of class 1 and class 0. Then the expected number of mislabeled observations (hypergeometric distribution) in each sample is rs/n_0, leading to a noise rate of

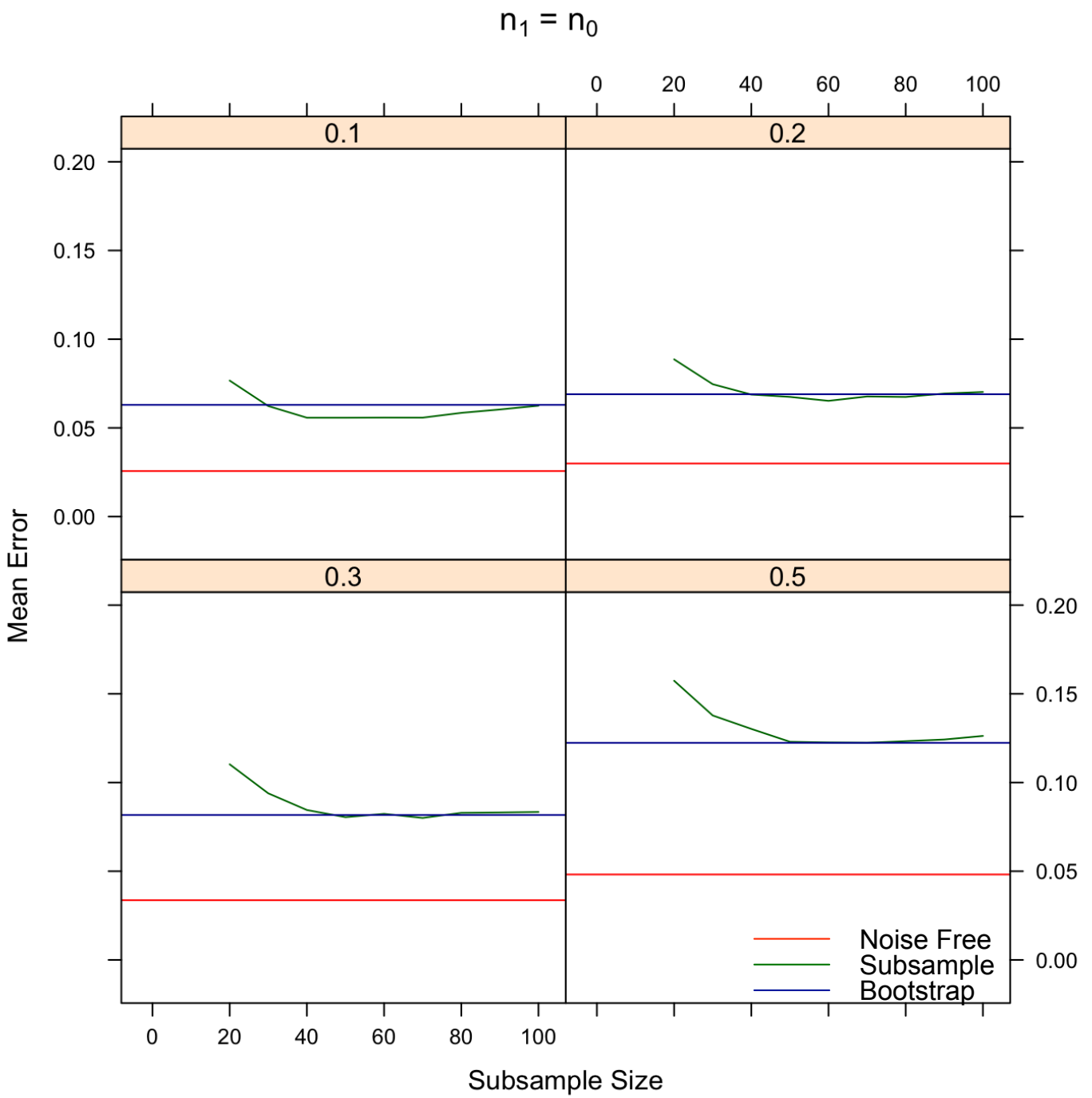
$$p_{noise} = r/2n_0$$

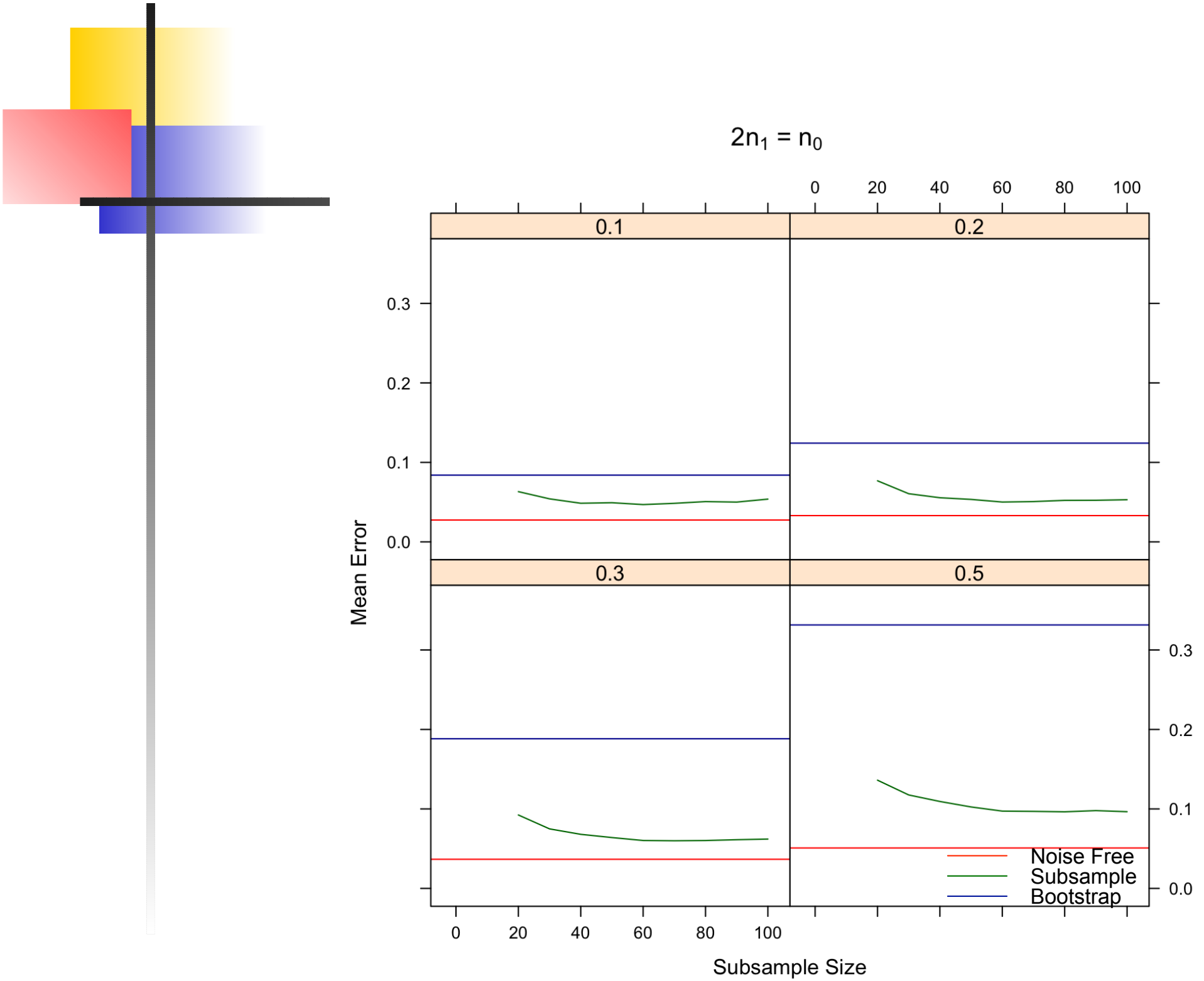
Sampling Strategy

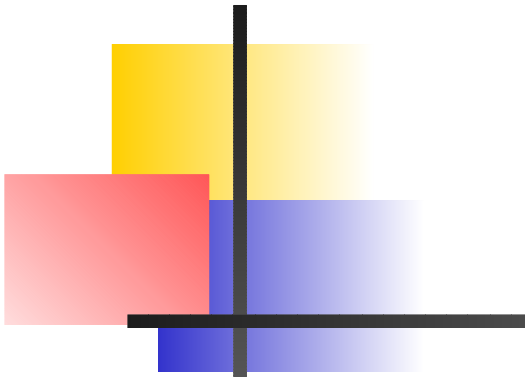
- Suppose our training sample contains $n = n_1 + n_0$ observations where n_1 are labeled as belonging to class 1 and n_0 are labeled as belonging to class 0.
- Suppose that r of the n_0 class 0 observations are mislabeled (they actually belong to class 1);
- Random Forest: the expected number of mislabeled observations in each bootstrap sample is r, leading to a constant noise rate of $p_{noise} = r/n$ for each tree in the forest.
- If we can reduce p_{noise} we may see performance improvements...
- Select (without replacement) $s < \min\{n_1, n_0\}$ observations from each of class 1 and class 0. Then the expected number of mislabeled observations (hypergeometric distribution) in each sample is rs/n_0, leading to a noise rate of

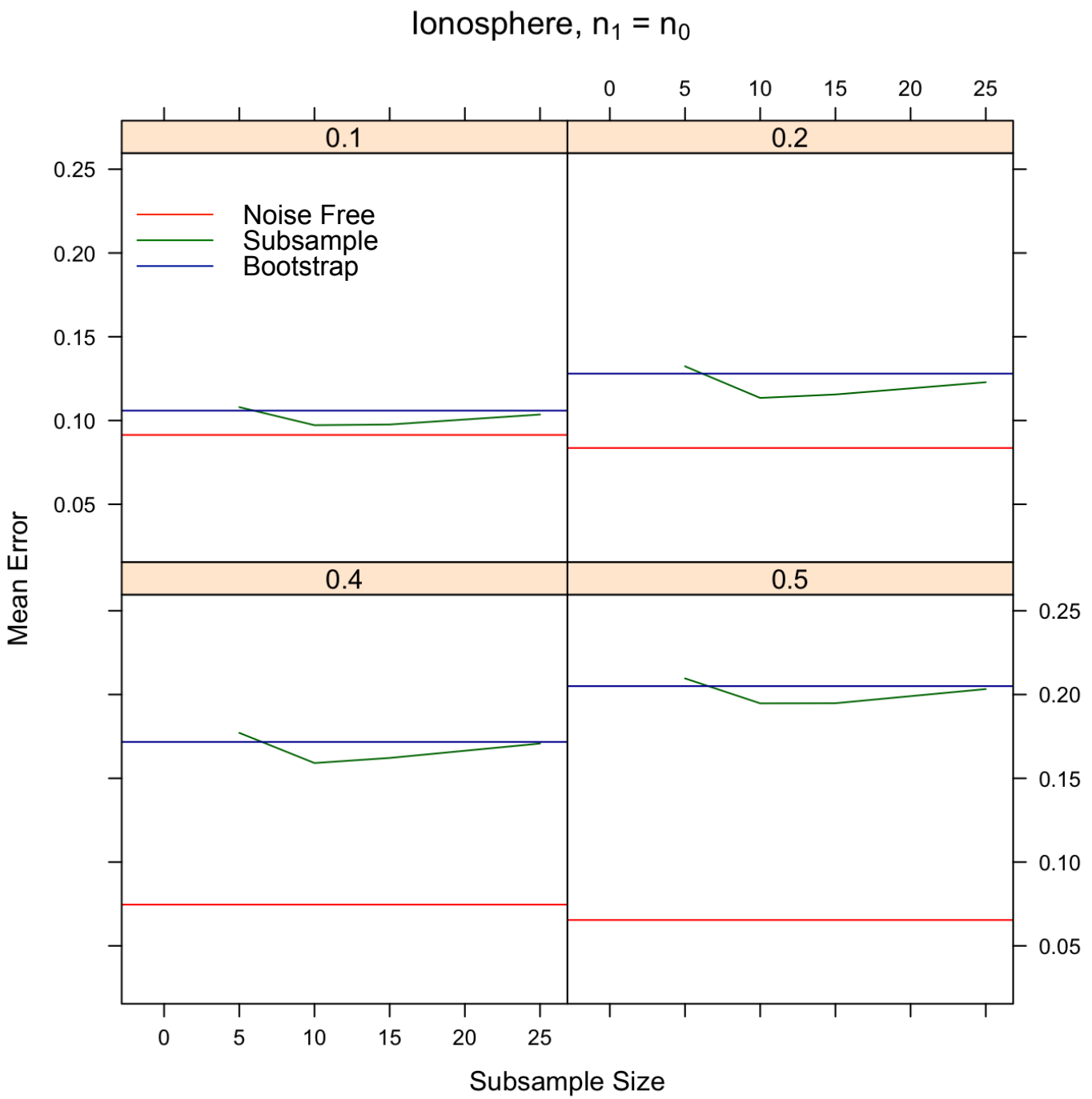
$$p_{noise} = r/2n_0$$

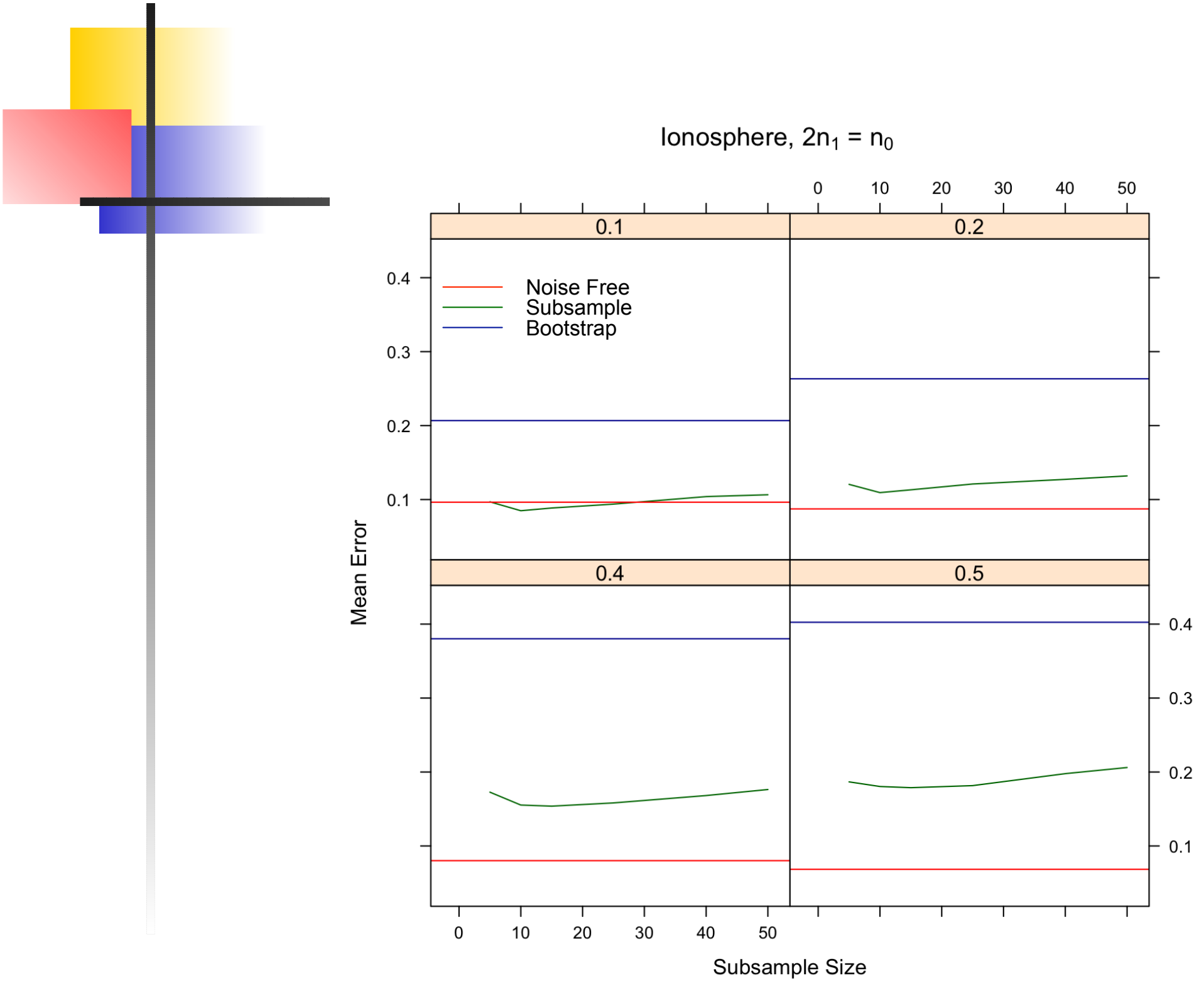
- Clearly, if $n_1 = n_0$, we should see no difference. However, if $n_1 < n_0$ we may see an improvement over bootstrap resampling.

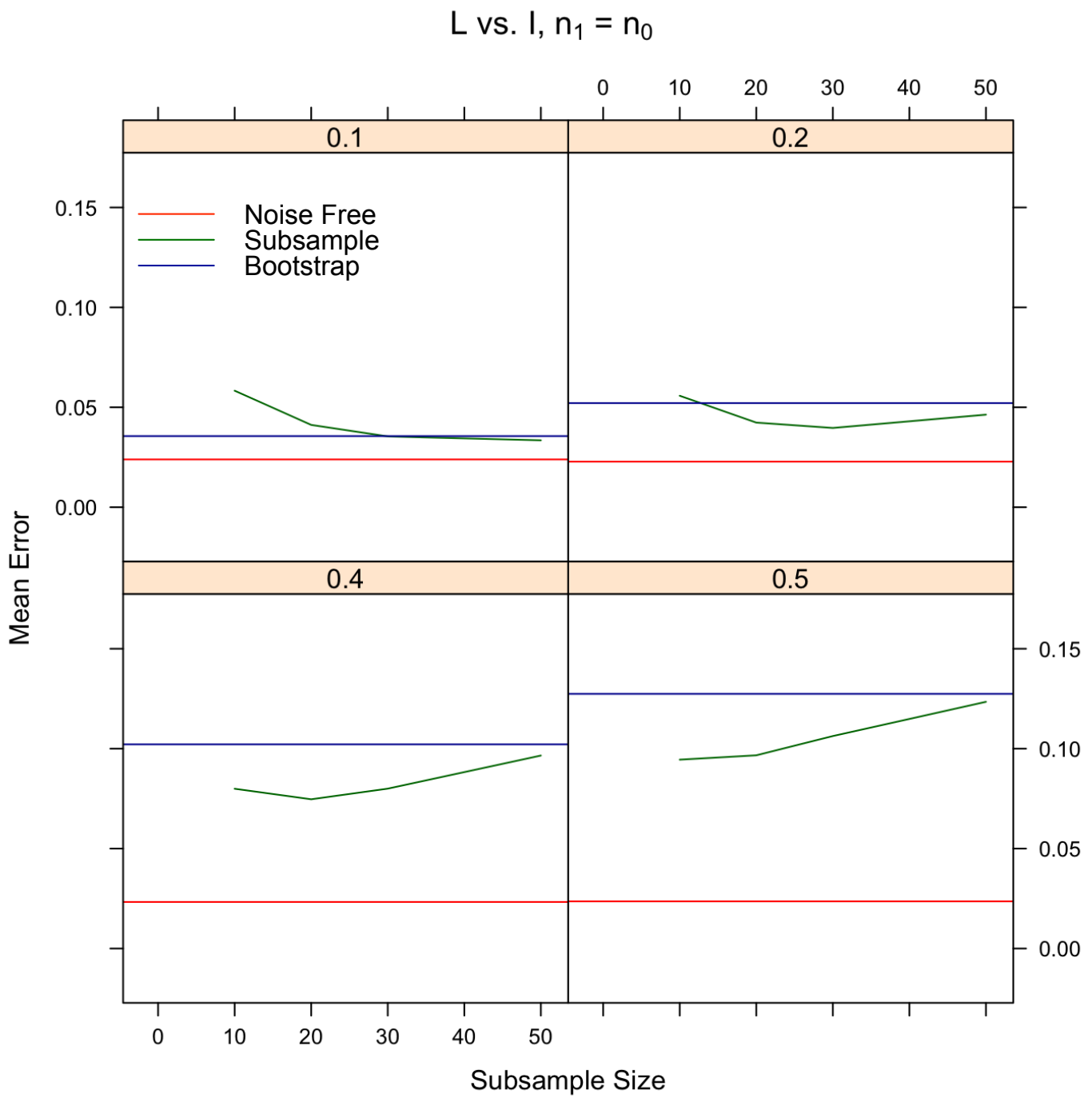

$n_1 = n_0$
0.1
0.2
0.3
0.5
Mean Error
Subsample Size
Noise Free
Subsample
Bootstrap

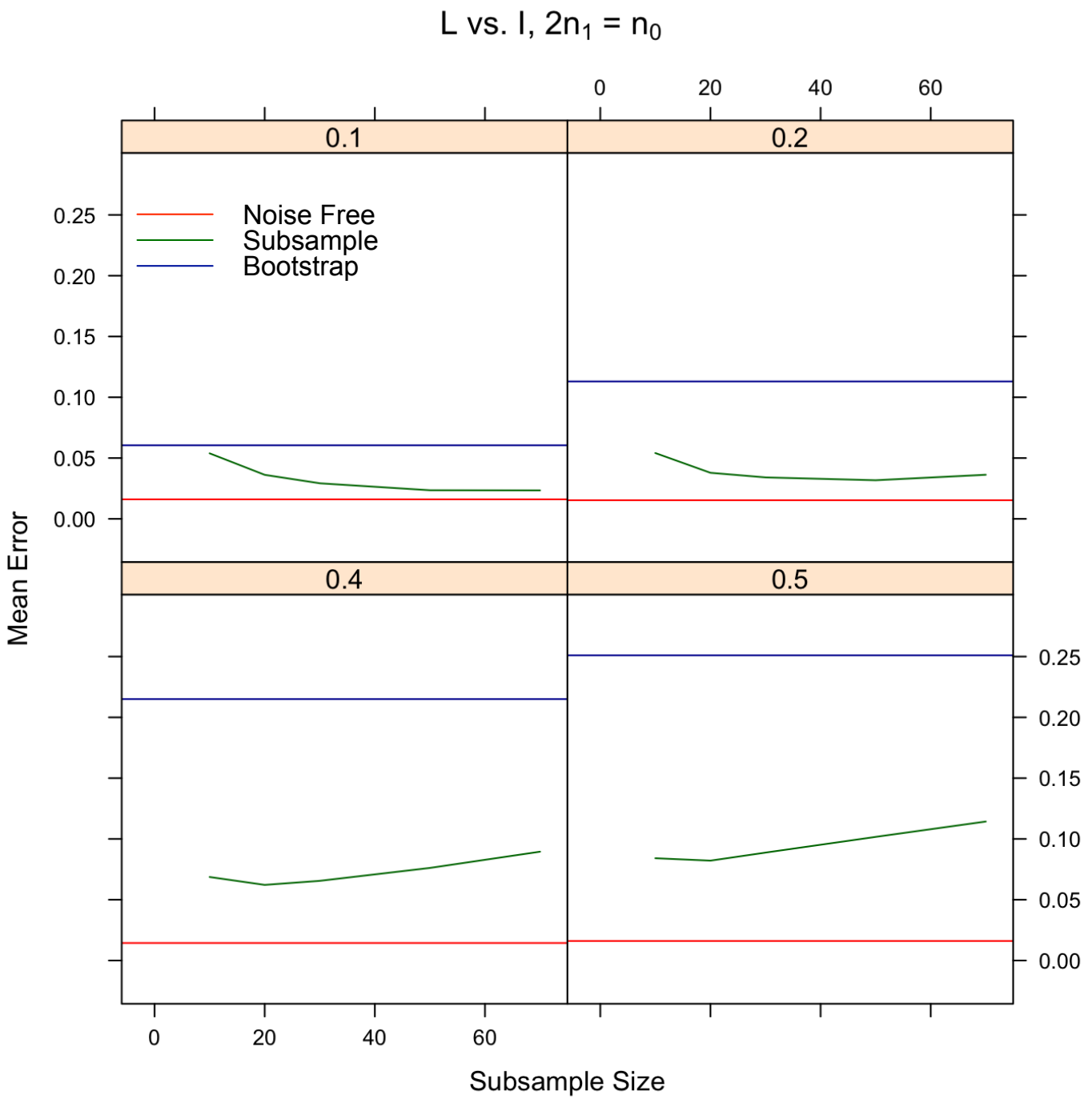

$2n_1 = n_0$
0.1
0.2
0.3
0.5
Mean Error
Subsample Size
Noise Free
Subsample
Bootstrap
0
20
40
60
80
100
0.0
0.1
0.2
0.3

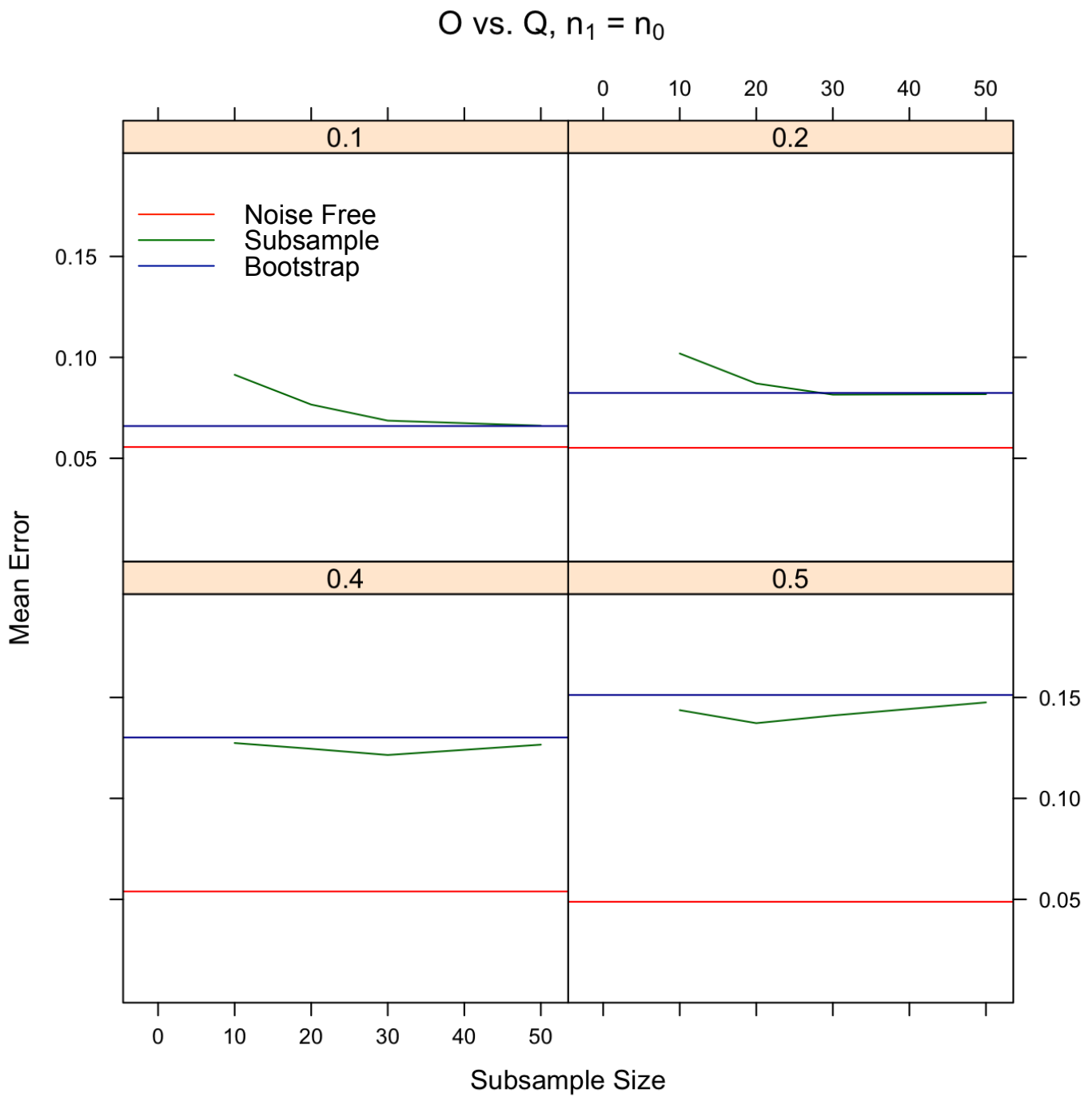


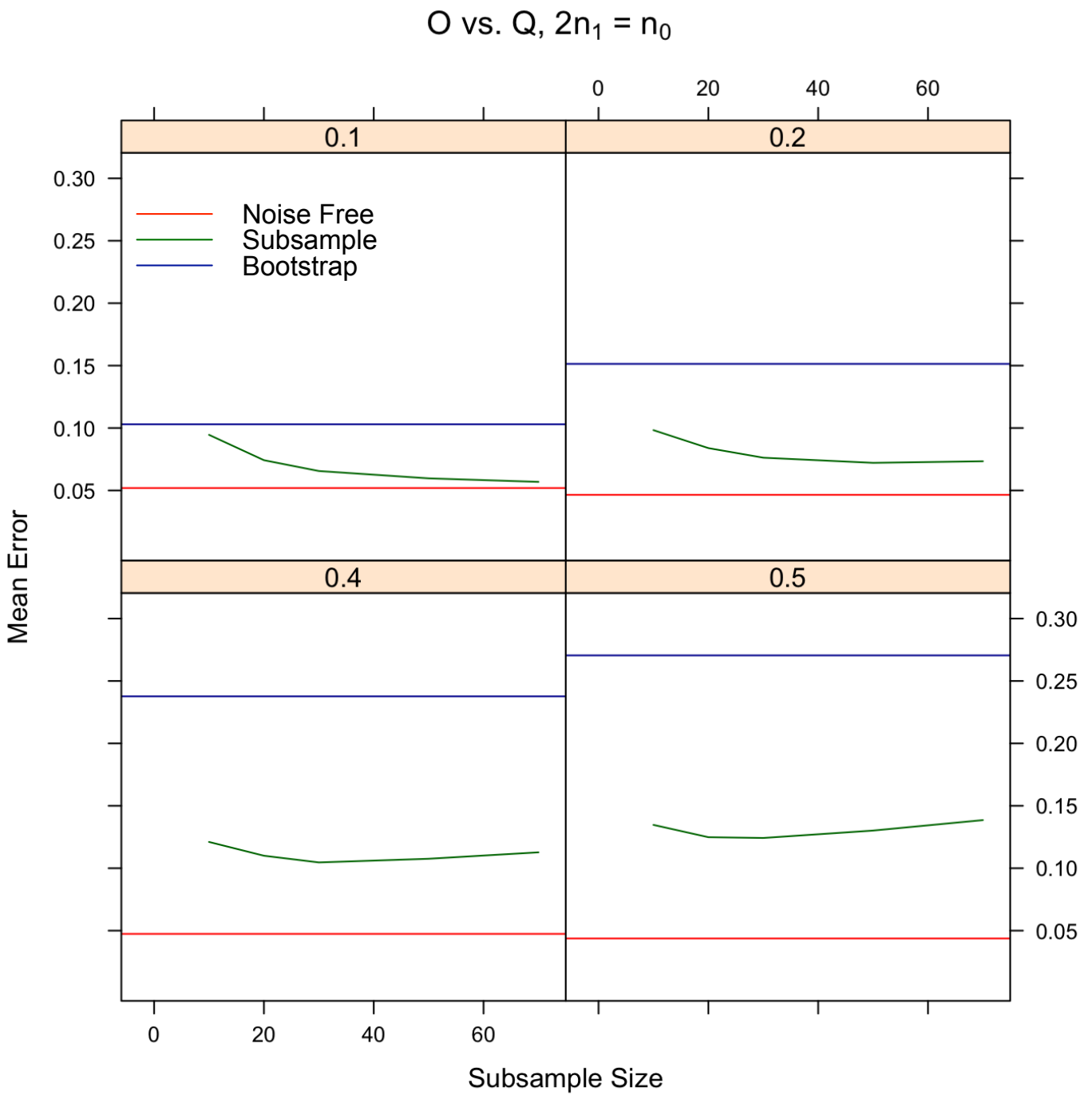

Ionosphere, $n_1 = n_0$
0.1
0.2
0.4
0.5
Noise Free
Subsample
Bootstrap
Mean Error
Subsample Size
0.25
0.20
0.15
0.10
0.05
0
5
10
15
20
25


Ionosphere, $2n_1 = n_0$
0.1
0.2
0.4
0.5
Noise Free
Subsample
Bootstrap
Mean Error
Subsample Size
0
10
20
30
40
50
0.1
0.2
0.3
0.4


L vs. l, $n_1 = n_0$
0.1
0.2
0.4
0.5
Noise Free
Subsample
Bootstrap
Mean Error
Subsample Size
0.00
0.05
0.10
0.15
0
10
20
30
40
50


L vs. l, $2n_1 = n_0$
0.1
0.2
0.4
0.5
Noise Free
Subsample
Bootstrap
Mean Error
Subsample Size
0.00
0.05
0.10
0.15
0.20
0.25
0
20
40
60


O vs. Q, $n_1 = n_0$
0.1
0.2
0.4
0.5
Noise Free
Subsample
Bootstrap
Mean Error
Subsample Size
0.05
0.10
0.15
0
10
20
30
40
50


O vs. Q, $2n_1 = n_0$
0.1
0.2
0.4
0.5
Noise Free
Subsample
Bootstrap
Mean Error
Subsample Size
0.05
0.10
0.15
0.20
0.25
0.30
0
20
40
60

Conclusions

- Two class classification problems with one sided noise

Conclusions

- Two class classification problems with one sided noise
- Have more observations from the noisy class than the non-noisy class...

Conclusions

- Two class classification problems with one sided noise
- Have more observations from the noisy class than the non-noisy class...
- Bootstrapping to construct tree ensembles is not recommended

Conclusions

- Two class classification problems with one sided noise
- Have more observations from the noisy class than the non-noisy class...
- Bootstrapping to construct tree ensembles is not recommended
- Instead, subsample each class separately