

CURSEDFOREST - A Random Forest Implementation for “Big” and “Wide” Data

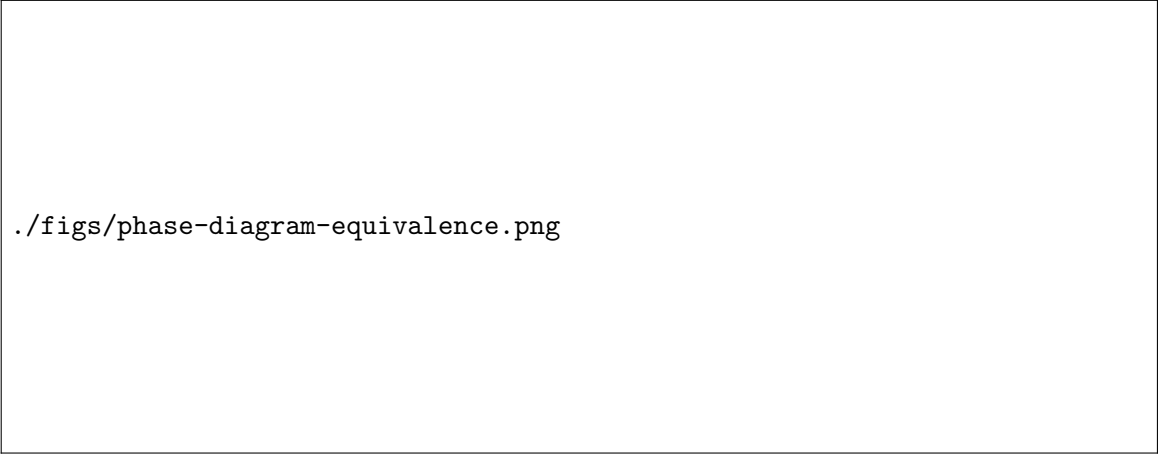
Supplementary Information

1 supplementary Information for Section 4.1

2 intro

We have demonstrated that using a different parallelization model can extend random forests to the case of an extremely large number of variables. We have treated the case of variable selection in a $p \gg n$ model where most of the variables are uninformative and have demonstrated the utility of the model for large GWAS datasets. By comparing this implementation to other implementations (including those optimized for large datasets) we have demonstrated the utility of this approach.

Donoho and Tanner ? give a “universal phase change” result that has applications in a large number of areas including variable selection in high dimensions. They show that there is a sharp disjunction between the cases where significant variables may be recovered with a high accuracy by procedures like stepwise variable selection, and the cases where they can not be recovered. The boundary is shown in Fig ?? in a space parameterized by the level of underdeterminedness, $\delta = n/p$, and by the sparsity, $\rho = k/n$ (where k is the number of significant variables). The theoretical boundary is based on arguments from combinatorial geometry. Above the phase-transition line variable recovery is still possible by a combinatorial approach such as all-subsets variable selection.



./figs/phase-diagram-equivalence.png

Figure 1: **The phase change diagram.**

2.1 the simulation

? investigate the behavior of a number of regression approaches for variable selection in a small

simulation. They consider a regression problem with $X_{n \times p} \sim N(0, 1)$ with p fixed at 200 and n variable. They set $\beta(1 : k) \sim U(1, 100)$ and $\beta((k + 1) : p) = 0$. Then $y = X\beta + \epsilon$, $\epsilon \sim N(0, \sqrt{16})$. They evaluate variable selection by the error measure

$$\frac{\|\hat{\beta} - \beta\|_2}{\|\beta\|_2}$$

They consider a number of variable selection methods, including a false discovery rate criteria. This involves adding the variable with the maximum t -value to the linear model if the p -value is less than $25(\text{number of terms currently in the model})/(\text{total number of variables})$. See Fig ?? for the error measure and figure ?? for the RBO, comparing the ranking (i.e. values) of $\hat{\beta}$ and β . It apparent that there is a marked drop in the accuracy of the variable recovery in line with the prediction of the Donoho-Tanner phase transition.

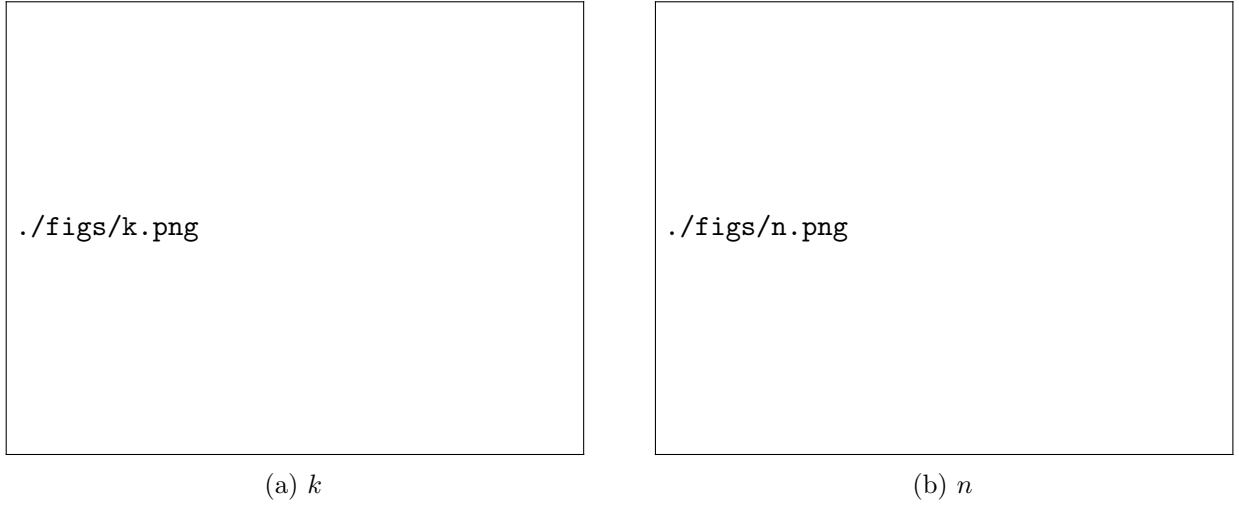


Figure 2: The simulation in ? considers $\delta = n/p$ and $\rho = k/n$. Here we plot the space of $\{\delta, \rho\}$, colored by the paramters n and k

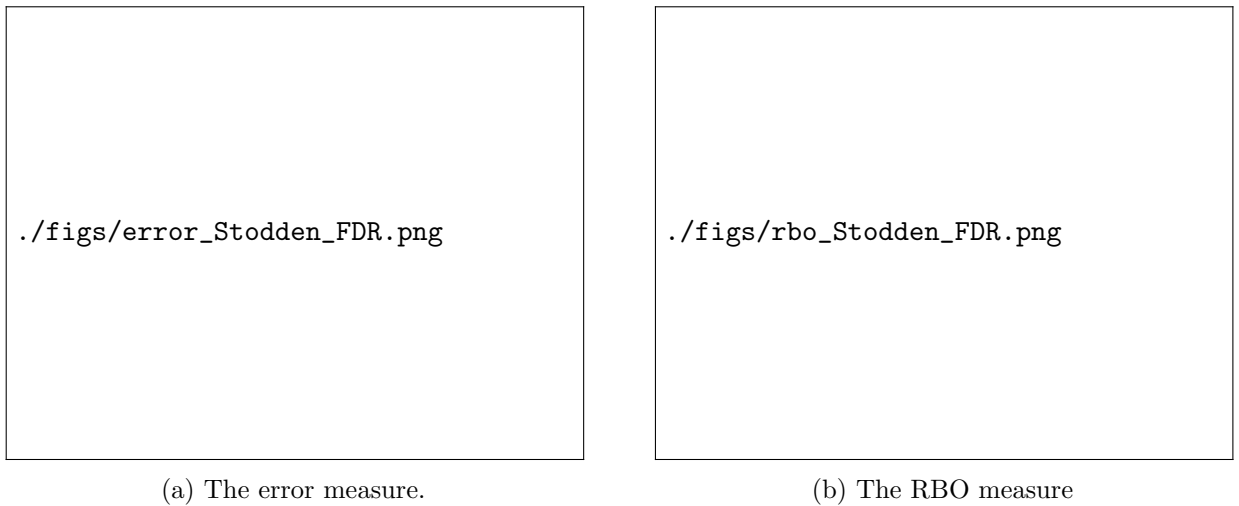


Figure 3: Forward stepwise variable selction with a false discovery rate stopping criteria.

2.2 RF on the simulation

We have seen no work on the question of such a phase transition in the case of classification problems. Still it amy be instructive to consider the implications of the pahse transition for our work. As CURSEDFORESTis designed for extremely large numbers of variables it is likely to be operating in



(a) Stepwise with FDR stopping criteria. $p = 200$, n and k variable.

(b) RBO

Figure 4

difficult regions of the figure where the ratio $\delta = n/p$ is small. In the simulated example in section ?? the underdeterminedness $\delta = 0.002$ and the sparsity $\rho = 2 \times 10^{-6}$

We have used a random forest on the same simulation as figure ?. The Donoho-Tanner phase transition arises in recovering the β in data generated by a linear model. However, in a decision tree (random forest) we are fitting a non-linear model and there is no notion of estimating the β . Because of this we have evaluated the performance of the random forest using the RBO measure on the variable importance.

We note that while a random forest is a long way from an all-subsets search, it is a limited combinatorial search of the feature space. As such it may perform outside of the bounds of the phase transition. combinatorial search.

Fig ?? shows the OOB prediction error and the RBO error for a random forest with 10000 trees and the mtry value set to the default for a regression problem (depending on n). It shows no evidence of following the shape of the Donoho-Tanner phase transition.

2.3 Conclusion

As CURSEDFOREST is designed for extremely large numbers of variables it is likely to be operating in difficult regions of the figure. We note that the examples we are considering lie on the extreme left of the plot, close to the origin. So there are limits to what can be recovered in the presence of noise. However in the case of data that is both big and wide, CURSEDFOREST and other VariantSpark methods may provide a useful tool.

The existence of the Donoho-Tanner phase transition is a salutary warning. There are likely to be limits, both computational and logical, to the recovery of signals from noisy data. CURSEDFOREST is a contribution to addressing the practical limits but the logical limits will still apply. However in the case of data that is both big and wide, CURSEDFOREST and other VariantSpark methods may provide a useful tool.