The Donoho-Tanner Phase Transition

1 Introduction

We consider the question of variable recovery in a sparse linear system $y = X\beta + \epsilon$ where only a small number k of the β variables are not equal to 0. We would like the solution,

$$\min_{\beta} \|\beta\|_0 \text{ s.t. } y = X\beta,$$

however; this is computationally unfeasible for larger numbers of variables. We have to approximate the solution by minimizing some l_1 or l_2 criteria, or by a more limited heuristic search of the solution space. Donoho and Tanner (2009) and Donoho and Stodden (2006) have considered this problem and introduce a "phase diagram" to help explore the behaviour of various approaches. This is implemented in the Matlab library `https://sparselab.stanford.edu`.

My aims in this exercise have been threefold:

- to investigate these ideas in R. To this end I have replicated some of the plots from Donoho and Stodden (2006);
- to see if a ranking method like rank-biased-overlap (RBO) (Webber et al., 2010) will allow us to extend some of these ideas to methods that do not produce an estimate of the β coefficients. See section 5 for a brief discussion of RBO;
- to see how a random forest behaves on the simulation used in Donoho and Stodden (2006).

2 The phase transition

Donoho and Tanner (2009) give a "universal phase change" result that has applications in a large number of areas, including variable selection in high dimensions. The theoretical phase change boundary is based on arguments from combinatorial geometry.

They argue that there is a sharp disjunction between the cases where informative 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 1 in the "phase space" parameterized by the level of underdetermination, $\delta = n/p$, and by the sparsity, $\rho = k/n$ (where k is the number of informative variables). Above the phase-transition line variable recovery is still possible by an exaustive search.

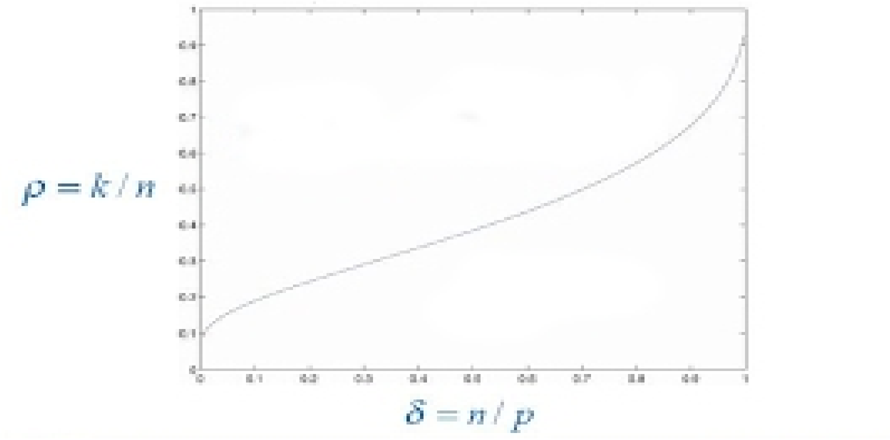


Figure 1: The phase change diagram. Below the line informative variables may be recovered with a high accuracy. Above the line a combinatorial search is required. Reproduced from Donoho and Stodden (2006)

3 A simulation

Donoho and Stodden (2006) investigate the phase transition in a small simulation. They consider a regression problem with $\underset{n\times p}{X} \sim N(0,1)$ with p fixed at 200 and n variable. Figures 2a and 2b show the phase space colored by n and k.

They set $\beta(1:k) \sim U(1,100)$ and $\beta((k+1):p) = 0$, then $y = X\beta + \epsilon$ with $\epsilon \sim N(0,\sqrt{16})$. They evaluate variable selection by the normalized l_2 error measure

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

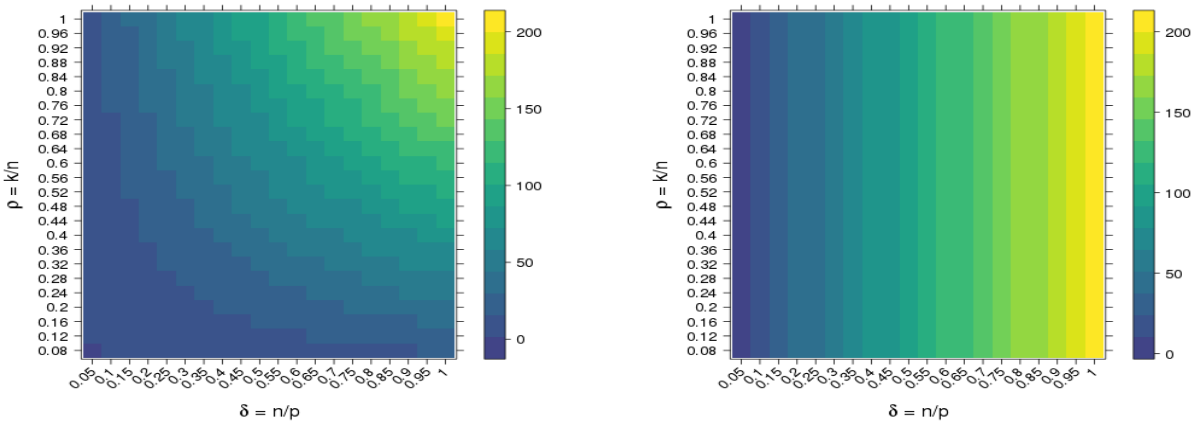
They consider a number of variable selection methods, including a false discovery rate criteria. This involves adding the variable with the maximum t-value to the linear model if the p-value is less than 0.25(number of terms currently in the model)/(total number of variables). See Fig 3a for the error measure and figure 3b for the RBO, comparing the ranking (i.e. values) of $\hat{\beta}$ and β. It apparent that the accuracy of the error measure shows a marked drop in line with the prediction of the Donoho-Tanner phase transition. The behaviour of the RBO measure is less clear.

4 Random Forests on the simulated data

We have used a random forest (the Ranger package, see Wright and Ziegler (2016)) on the same data as used for Fig 3. 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. Becsaue of this we have evaluted 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 search of the feature space. As such it may perform outside of the bounds of the phase transition.

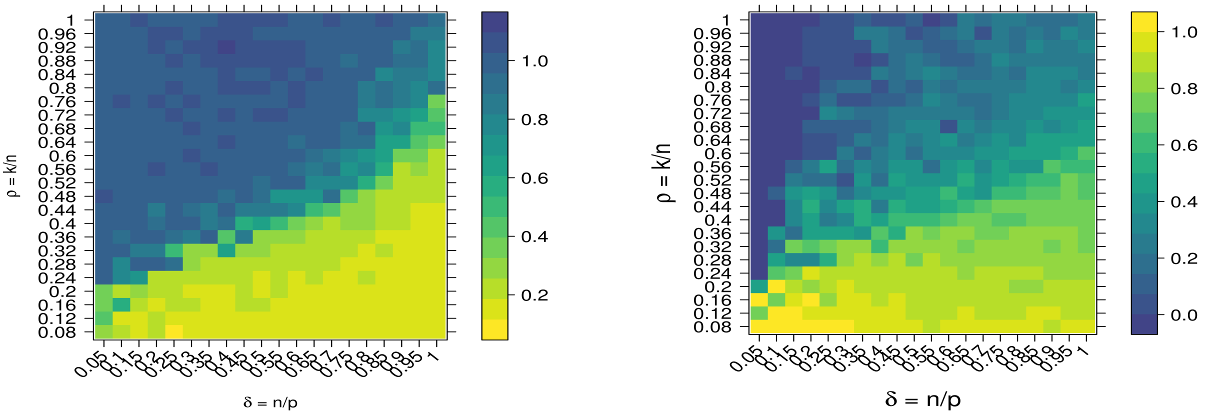
Fig 4 shows the OOB prediction error and the RBO error for a random forest with 10000 trees and the `mtry` value set to the default for a regression problem (depending on n). It shows no evidence of following the shape of the Donoho-Tanner phase transition.



(a) k

(b) n

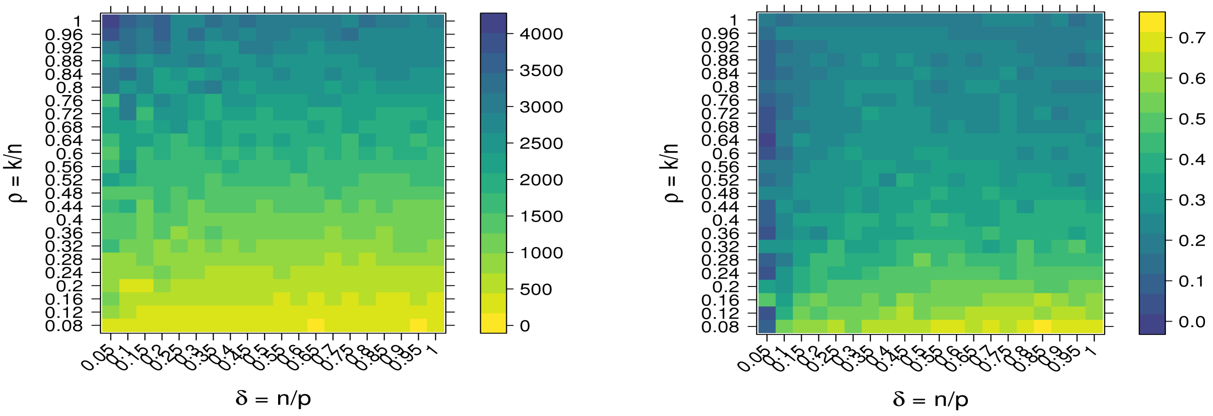
Figure 2: The simulation in Donoho and Stodden (2006) considers $\delta = n/p$ and $\rho = k/n$. Here we plot the space of $\{\delta, \rho\}$, colored by the paramters n and k.



(a) The error measure.

(b) The RBO measure

Figure 3: Forward stepwise variable selection with a false discovery rate stopping criteria. Note that the plot colors have been chosen so that the lighter color indicates the better result, that is, a lower error or a higher RBO.



(a) The out-of-bag error.

(b) The RBO measure.

Figure 4: The OOB error and RBO measure using a random forest.

5 the Rank-biased Overlap measure

Webber et al. (2010) Schmich (2015)

6 Conclusion

We have investigated the Donoho-Tanner phase transition in a small simulation, replicating some of the work of Donoho and Stodden (2006). We have investigated the use of the RBO measure for comparing the variable importance ranking produce by a random forest and the β parameters in a linear model used to define the data.

References

Donoho, D. and Stodden, V. (2006). Breakdown point of model selection when the number of variables exceeds the number of observations. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1916–1921. IEEE.

Donoho, D. and Tanner, J. (2009). Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906):4273–4293.

Schmich, F. (2015). *gespeR: Gene-Specific Phenotype EstimatoR*. R package version 1.1.2.

Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4):20:1–20:38.

Wright, M. N. and Ziegler, A. (2016). Ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software.* in press.