Performance of COR model in UCI data sets

Paula Parpart
Department of Experimental Psychology, University College London

October 15, 2015

1 Simulation

The model was fit to data sets from the UCI Machine Learning Repository (http://archive.ics.uci.edu/ml).

The chosen data sets are classification problems with a small number of continuous attributes and binary dependent variable. The COR model is cross-validated on each data set by splitting the total number of instances randomly into training and test set. The size of the training set is varied between 10, 20, 30, 50, 100 and 500 of all instances, and the test set represents the complementary set of instances always. For each training set size, the cross-validation split into training and test set is repeated 10 times and performance is averaged across all of them. Plots below demonstrate the generalization performance of the COR model for a range of penalization parameters $\theta = [0, 700]$, and as a function of the training set size.
In the limit, even when convergence is not quite reached, that is, the off-diagonal-weights $\neq 0$ still, I did not threshold them to zero artificially.

The decision rule applied here does not reflect the heuristic decision rules (TTB or Tallying) , but a standard linear classification rule which sums the outputs and thresholds at zero (i.e., like Linear Regression classification).

$$D\text{: Number of dimensions (i.e. cues)}$$

$$\hat{y} = 1\left[0 < \textstyle\sum_{i=1}^{D} x\hat{\beta}^{(i)} + \hat{\beta}_0^{(i)}\right]$$

1.1 Banknotes

Number of instances: 1372
Number of attributes: 4
Attribute Information:
1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (integer) : genuine versus forged

Description: Data were extracted from images that were taken for the evaluation of an authentication procedure for banknotes.

Average (absolute) correlation between attributes: 0.43
Minimum (absolute) correlation between attributes: 0.26
Maximum (absolute) correlation between attributes: 0.79

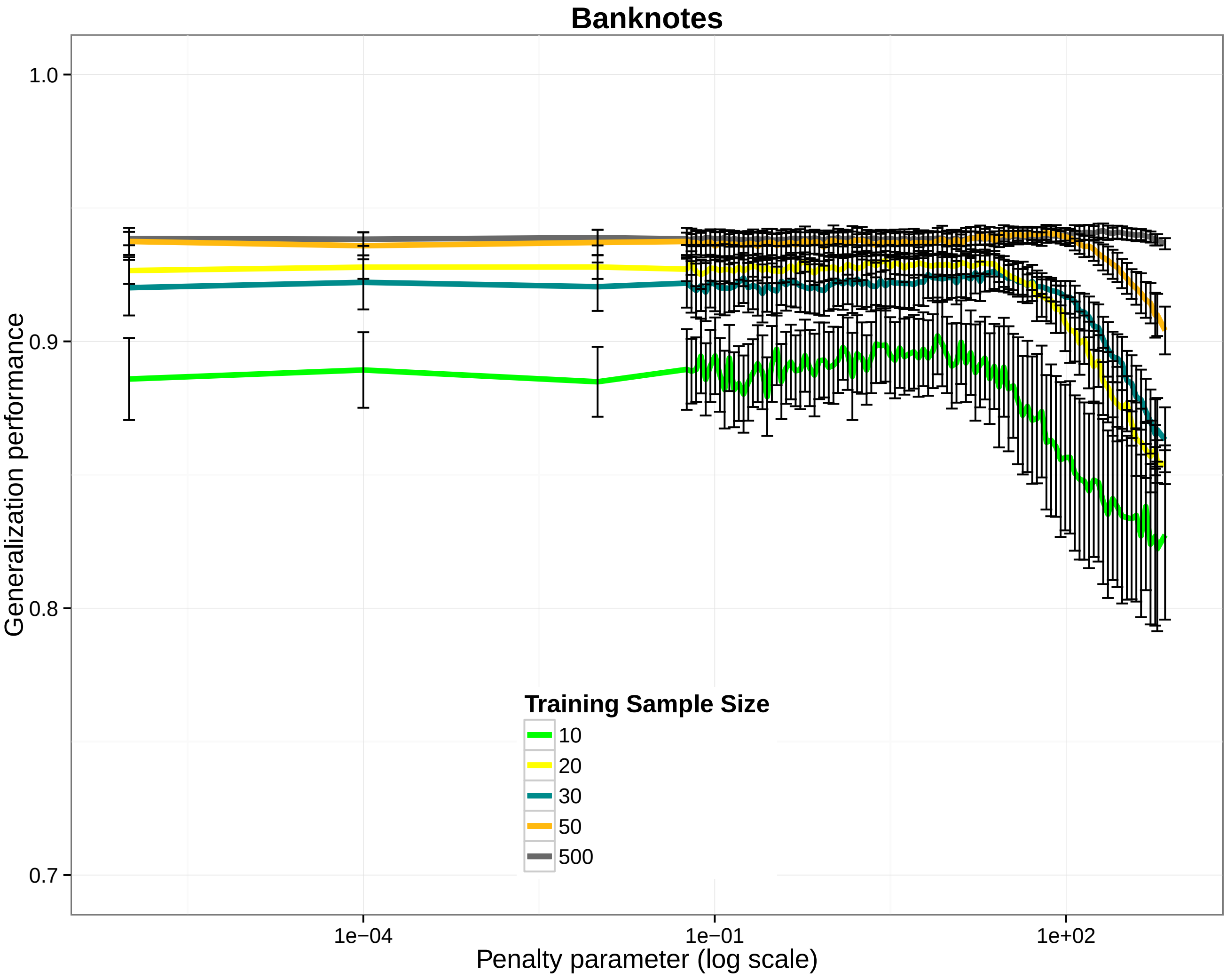

Figure 1: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, 100 and 500 of all instances.

1.2 Banknotes (Incl. Interactions)

Number of instances: 1372
Number of attributes: 10
Attribute Information:
Same 4 attributes as above, as well as all possible interactions between 4 Predictors. class (integer) : genuine versus forged

Average (absolute) correlation between attributes: 0.34
Minimum (absolute) correlation between attributes: 0.002
Maximum (absolute) correlation between attributes: 0.79

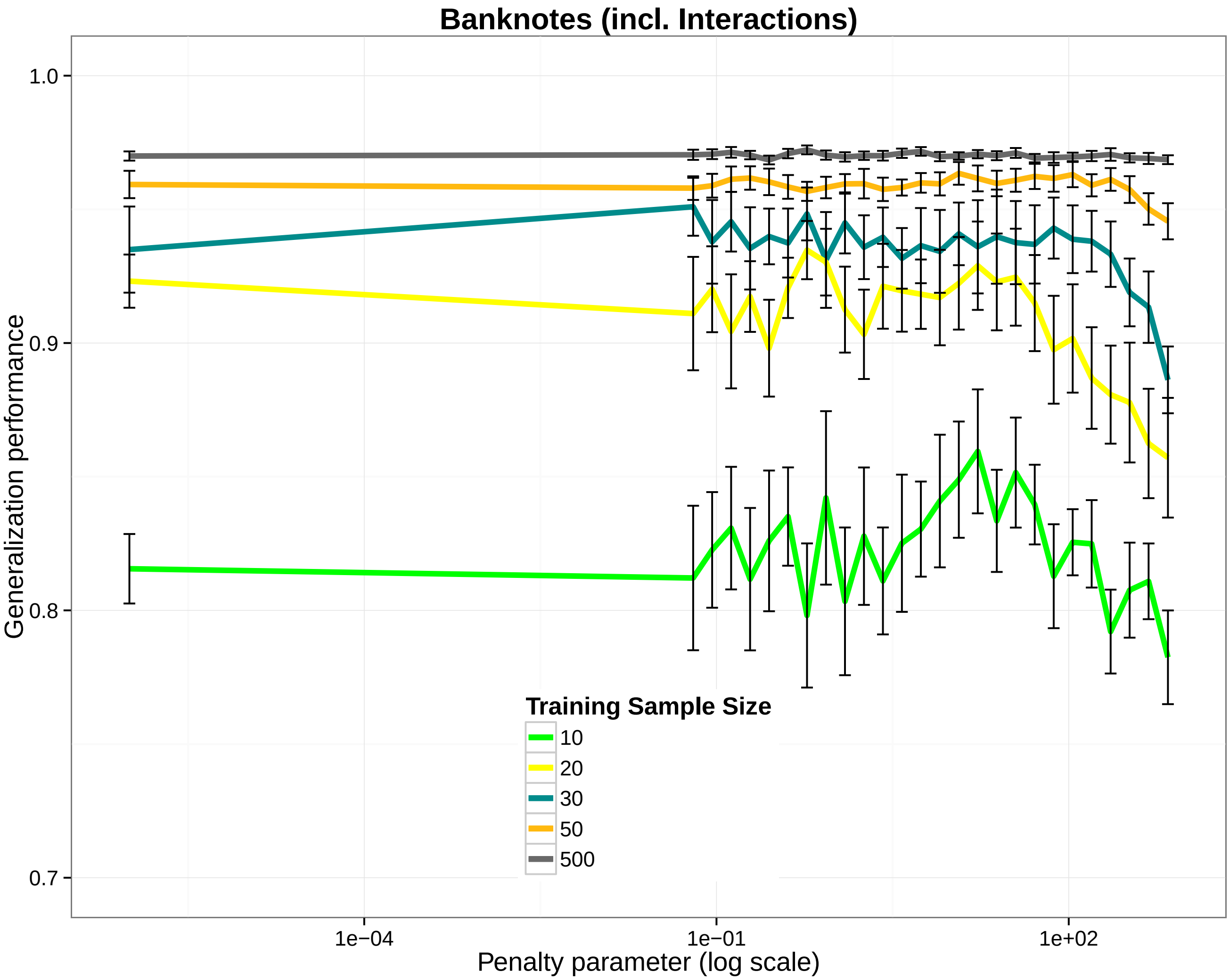

Figure 2: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, 100 and 500 of all instances.

1.3 Diabetes

Number of instances: 768
Number of attributes: 8
Attribute Information:
1. Number of times pregnant
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/(height in m)^2)
7. Diabetes pedigree function
8. Age (years)
9. Class variable (0 or 1) (class value 1 is interpreted as ”tested positive for diabetes”)
Average (absolute) corr: 0.17. Minimum (absolute) corr: 0.01. Maximum (absolute) corr: 0.54

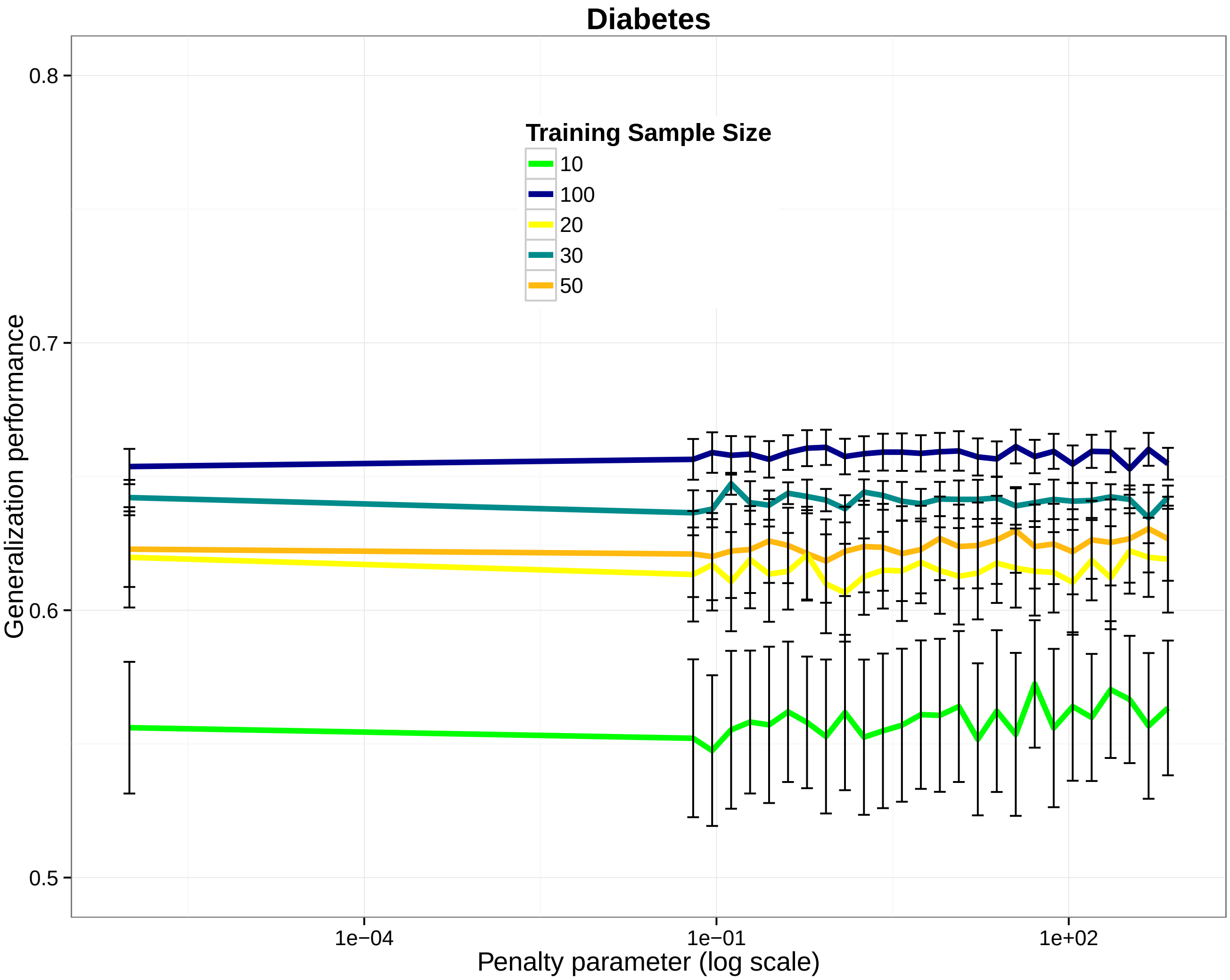

Figure 3: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, and 100 of all instances.

1.4 Haberman's Survival Data

Number of instances: 306
Number of attributes: 3
Attribute Information:
1. Age of patient at time of operation (numerical)
2. Patient's year of operation (year - 1900, numerical)
3. Number of positive axillary nodes detected (numerical)
4. Survival status (class attribute)
– 1 = the patient survived 5 years or longer
– 2 = the patient died within 5 year
Description: The dataset contains cases from a study that was conducted between 1958 and 1970 at the University of Chicago's Billings Hospital on the survival of patients who had undergone

surgery for breast cancer. Weights are not very large at penalty = 0, i.e. Linear Regression.

Average (absolute) correlation between attributes: 0.05
Minimum (absolute) correlation between attributes: 0.003
Maximum (absolute) correlation between attributes: 0.089

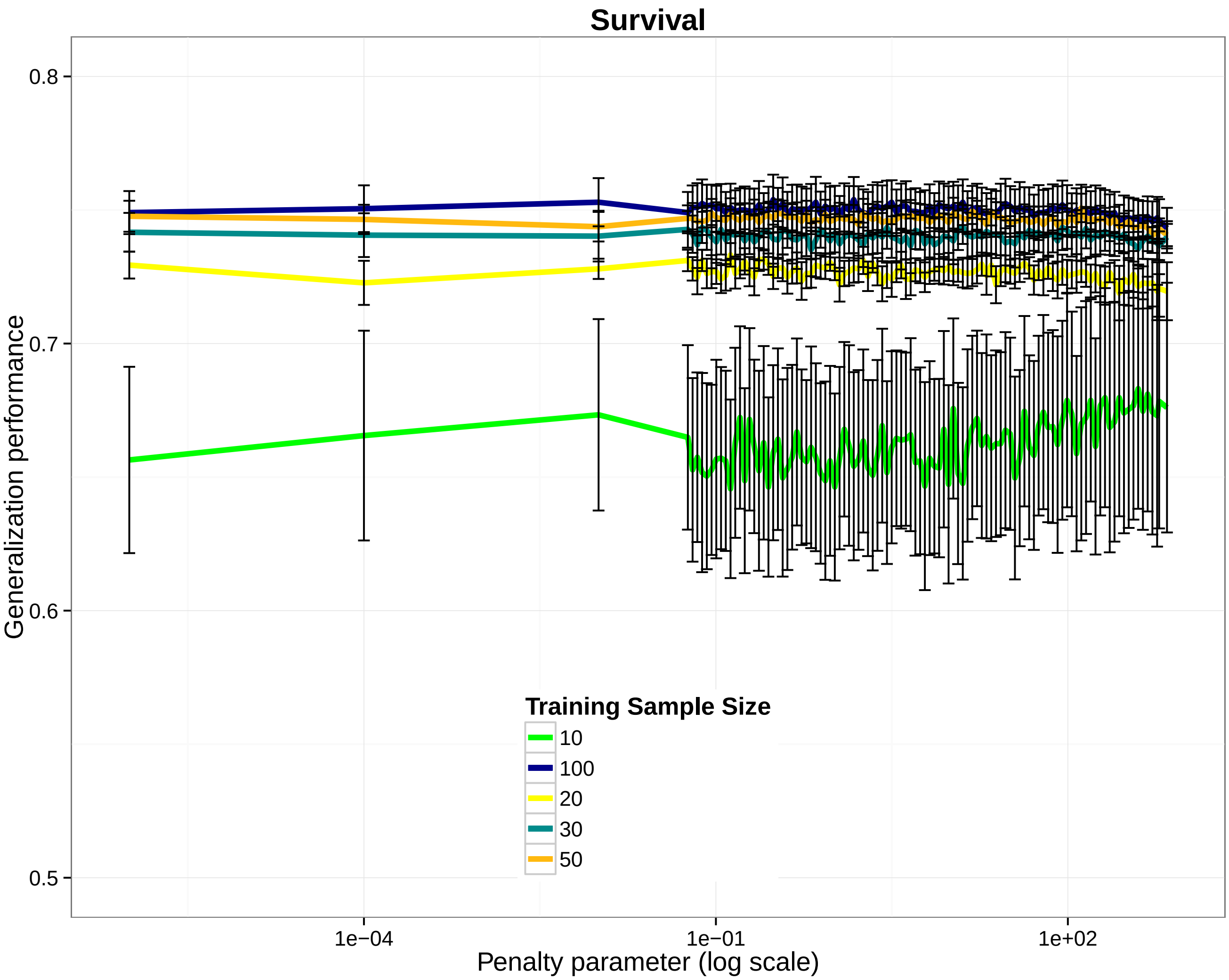

Figure 4: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, and 100 of all instances.

1.5 Iris (without Virginica)

Number of instances: 100 (without Virginica)
Number of attributes: 4
Attribute Information:
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
– Iris Setosa
– Iris Versicolour
– Iris Virginica
Description: This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

I deleted the 3rd class Virginica so the problem becomes a binary classification problem. I am repeating the analysis with a different classes removed further below.

Average (absolute) correlation between attributes: 0.66
Minimum (absolute) correlation between attributes: 0.21
Maximum (absolute) correlation between attributes: 0.98

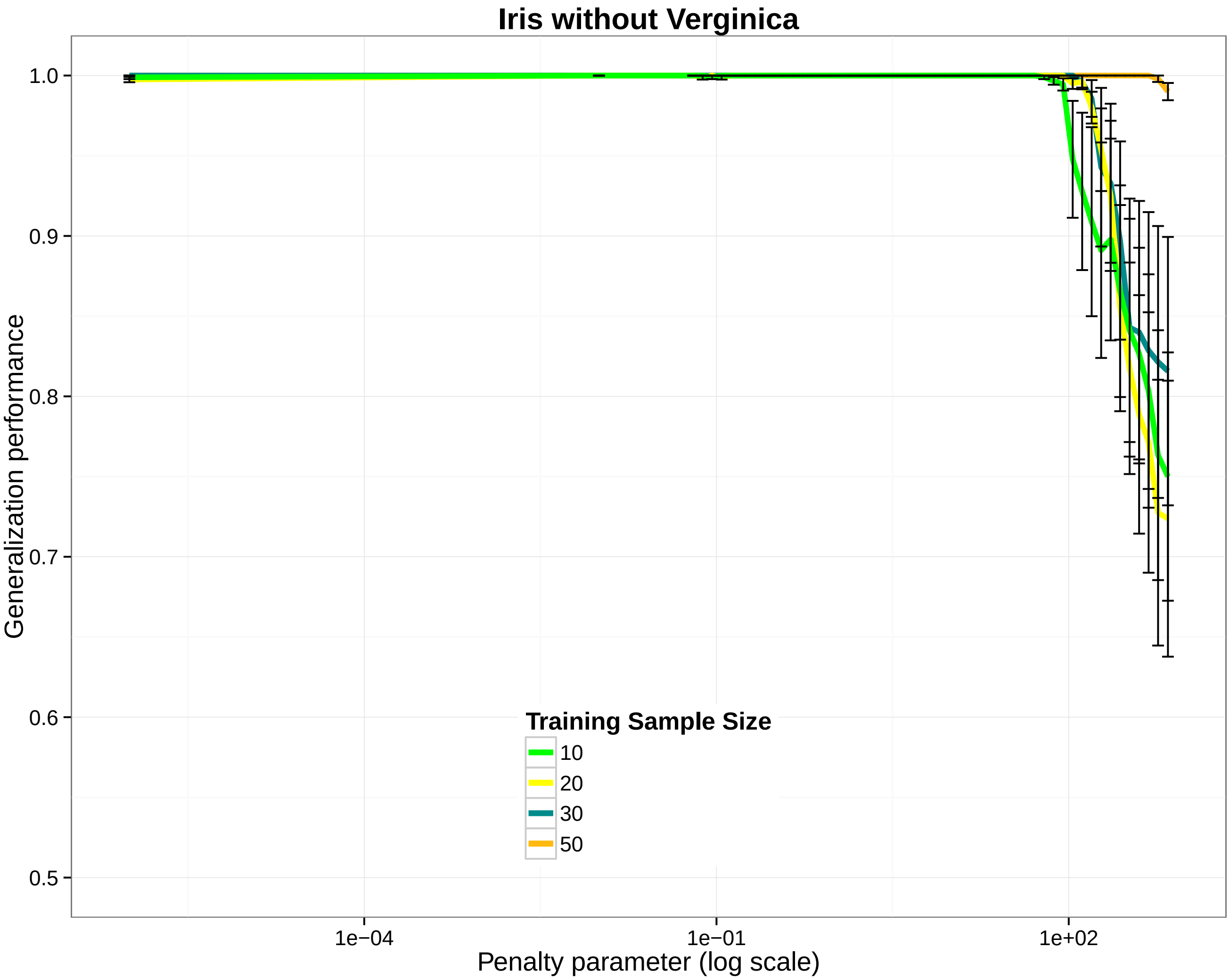

Figure 5: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, and 50 of all instances.

1.6 Iris (without Versicolor)

Number of instances: 100 (without Versicolor)
Number of attributes: 4
Attribute Information:
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class:
– Iris Setosa
– Iris Versicolor
– Iris Virginica

I deleted the class Versicolour so the problem becomes a binary classification problem. The classification performance is equally good as above (with Virginica removed), most likely because I compared the class ”Setosa” with either Versicolour or Virginica which leads to perfect classification performance. This should be lower when Versicolour and Virginica are compared as these 2 classes are not linearly separable.

Average (absolute) correlation between attributes: 0.65
Minimum (absolute) correlation between attributes: 0.22
Maximum (absolute) correlation between attributes: 0.97

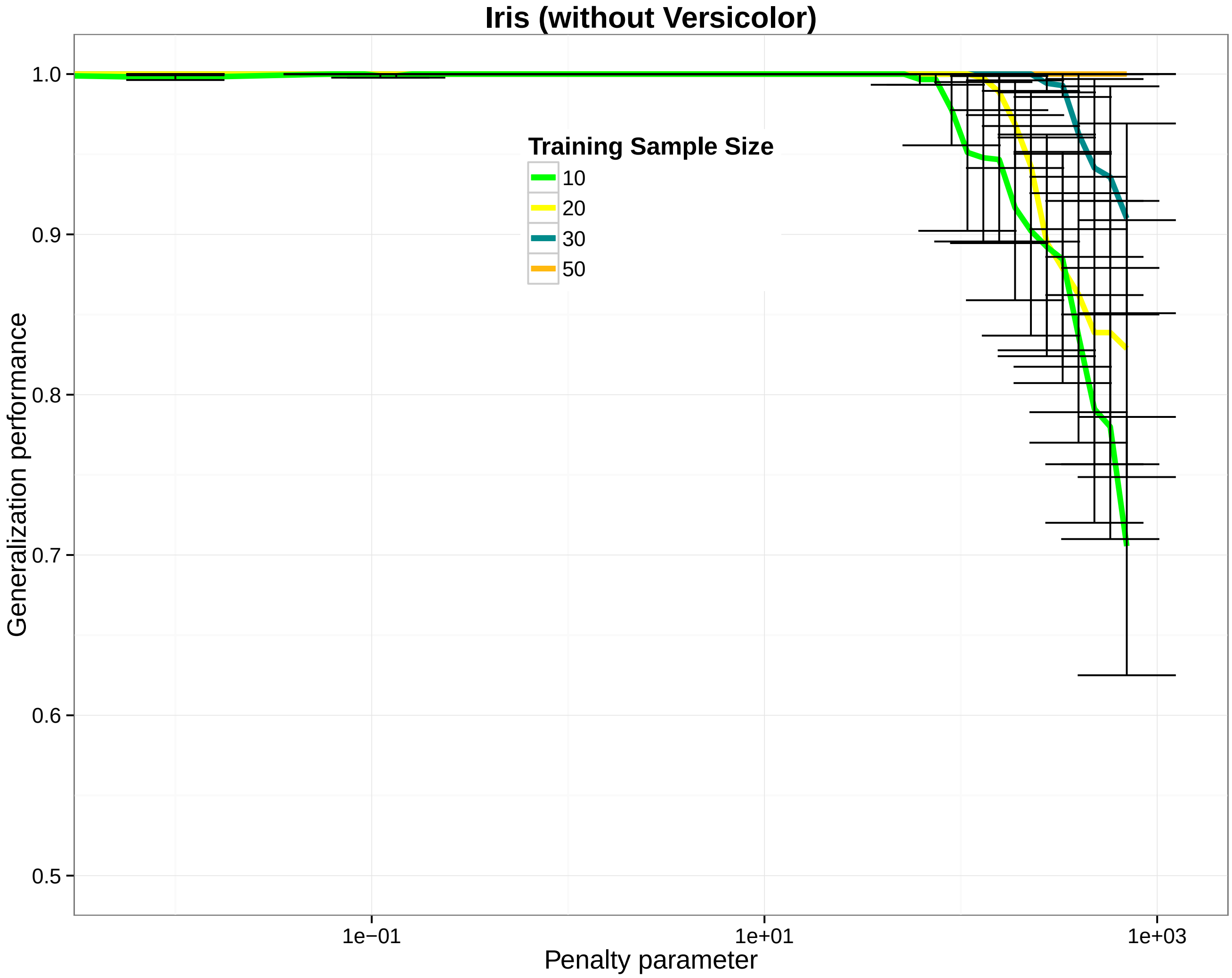

Figure 6: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, and 50 of all instances.

1.7 Iris (without Setosa)

Number of instances: 100 (without Setosa)
Number of attributes: 4
Attribute Information:
1. sepal length in cm
2. sepal width in cm
3. petal length in cm
4. petal width in cm
5. class: – Iris Setosa – Iris Versicolor – Iris Virginica
I deleted the class Setosa so the problem becomes a binary classification problem.

Average (absolute) corr: 0.65. Min (absolute) corr: 0.52. Max (absolute) corr: 0.82

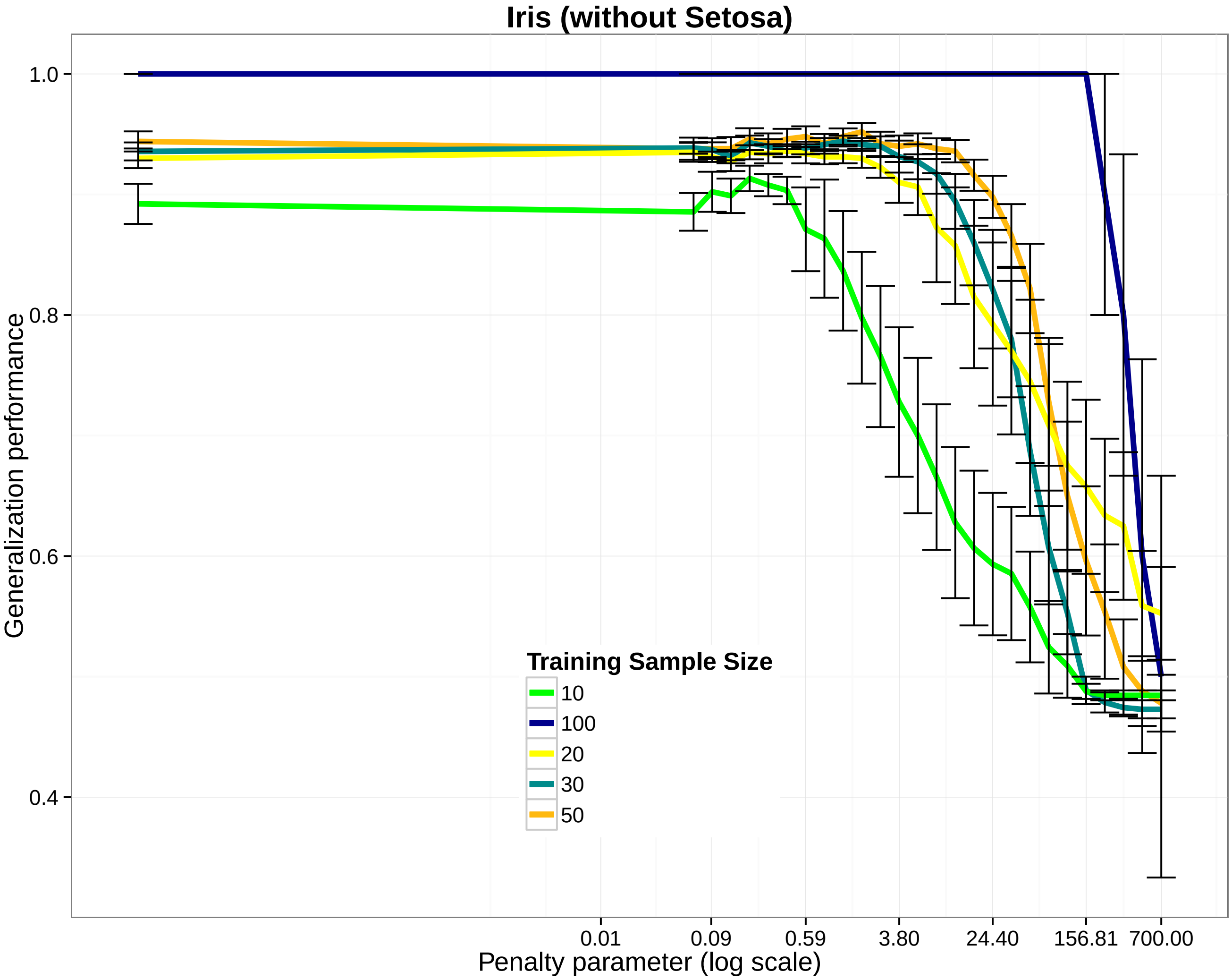

Figure 7: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, and 50 of all instances.

1.8 Skincolour

Number of instances: 245057
Number of attributes: 3
Attribute Information:
B, G, R (x1,x2, and x3 features) values representing color space,
class label (decision variable y): skin sample versus non-skin sample.
Skin and Nonskin dataset is generated using skin textures from face images of diversity of age, gender, and race people. Average (abs) corr: 0.67. Min (abs) corr: 0.50. Max (abs) corr: 0.86

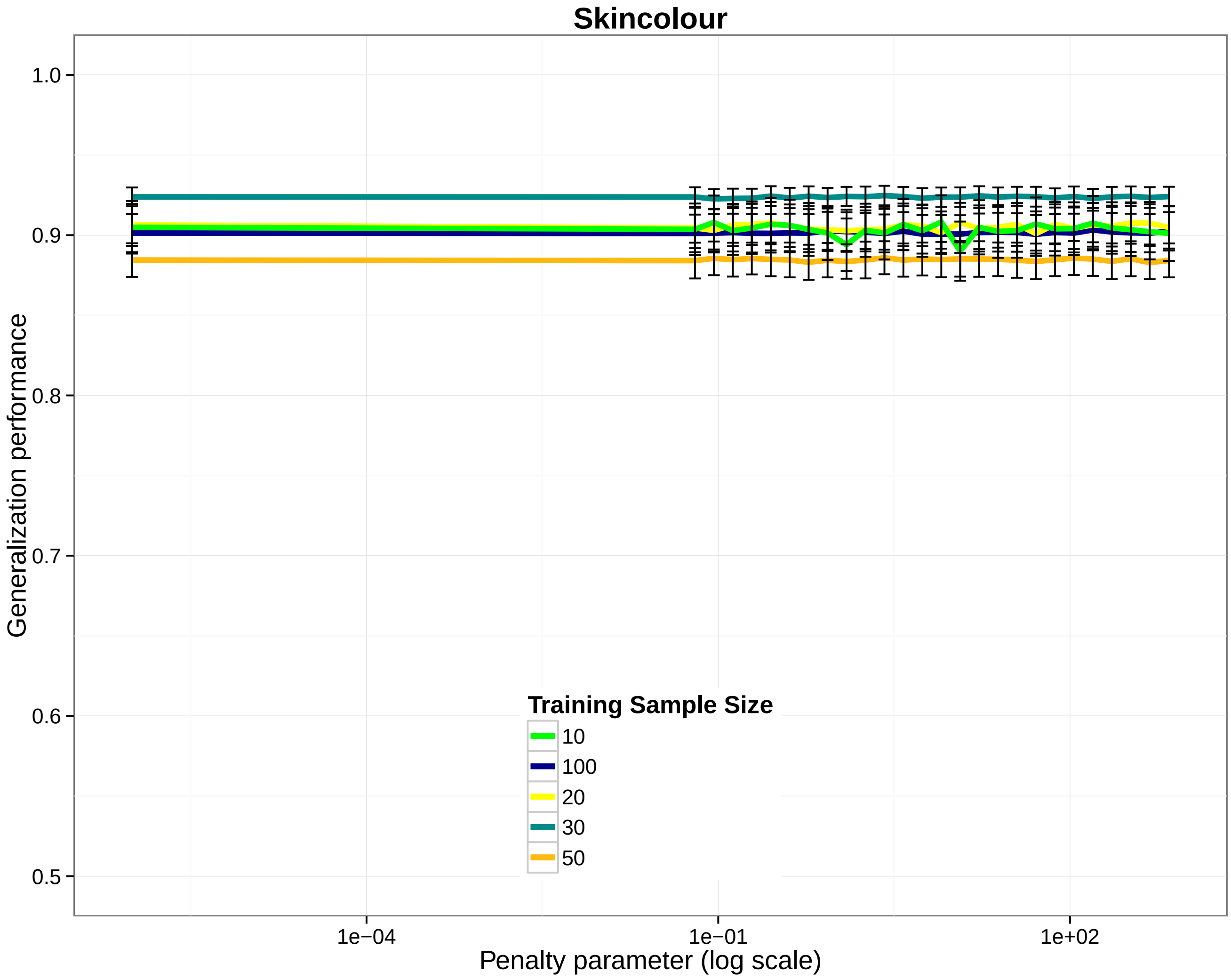

Figure 8: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, 100 of all instances.

1.9 Titanic

Number of instances: 2201
Number of attributes: 3
Attribute Information:
1. Class 2.Age 3.Sex
class label: Survived (yes or no)

Average (abs) corr: 0.19. Min (abs) corr: 0.07. Max (abs) corr: 0.38

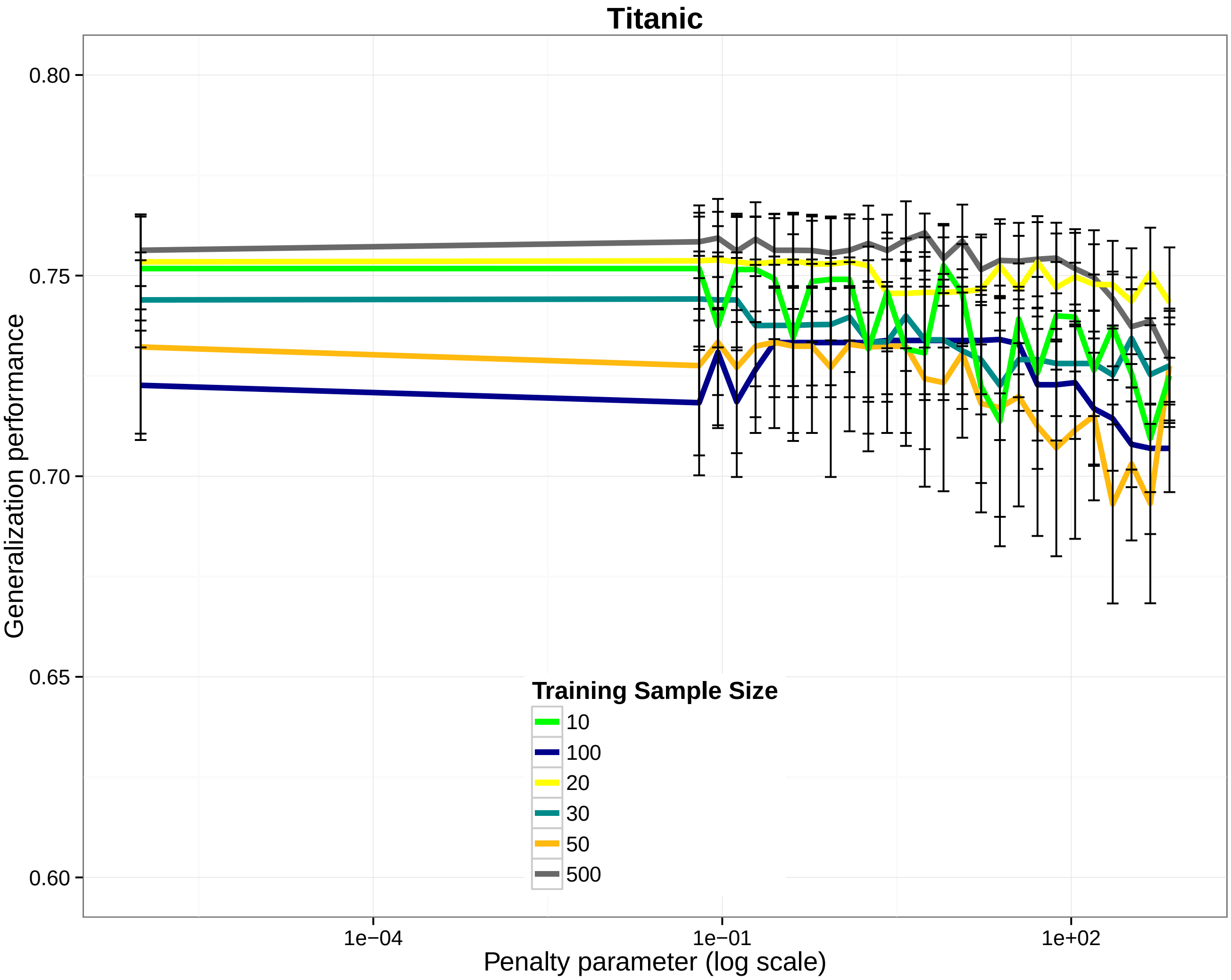

Figure 9: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, 100, and 500 of all instances.

1.10 TicTacToe

Number of instances: 958
Number of attributes: 9
Attribute Information:
1. top-left-square: x,o,b 2. top-middle-square: x,o,b 3. top-right-square: x,o,b 4. middle-left-square: x,o,b 5. middle-middle-square: x,o,b 6. middle-right-square: x,o,b 7. bottom-left-square: x,o,b 8. bottom-middle-square: x,o,b 9. bottom-right-square: x,o,b 10. Class: positive,negative
This database encodes the complete set of possible board configurations at the end of tic-tac-toe games, where ”x” is assumed to have played first. The target concept is ”win for x” (i.e., true when ”x” has one of 8 possible ways to create a ”three-in-a-row”). Average (abs) corr: 0.13. Min (abs) corr: 0.03. Max (abs) corr: 0.26

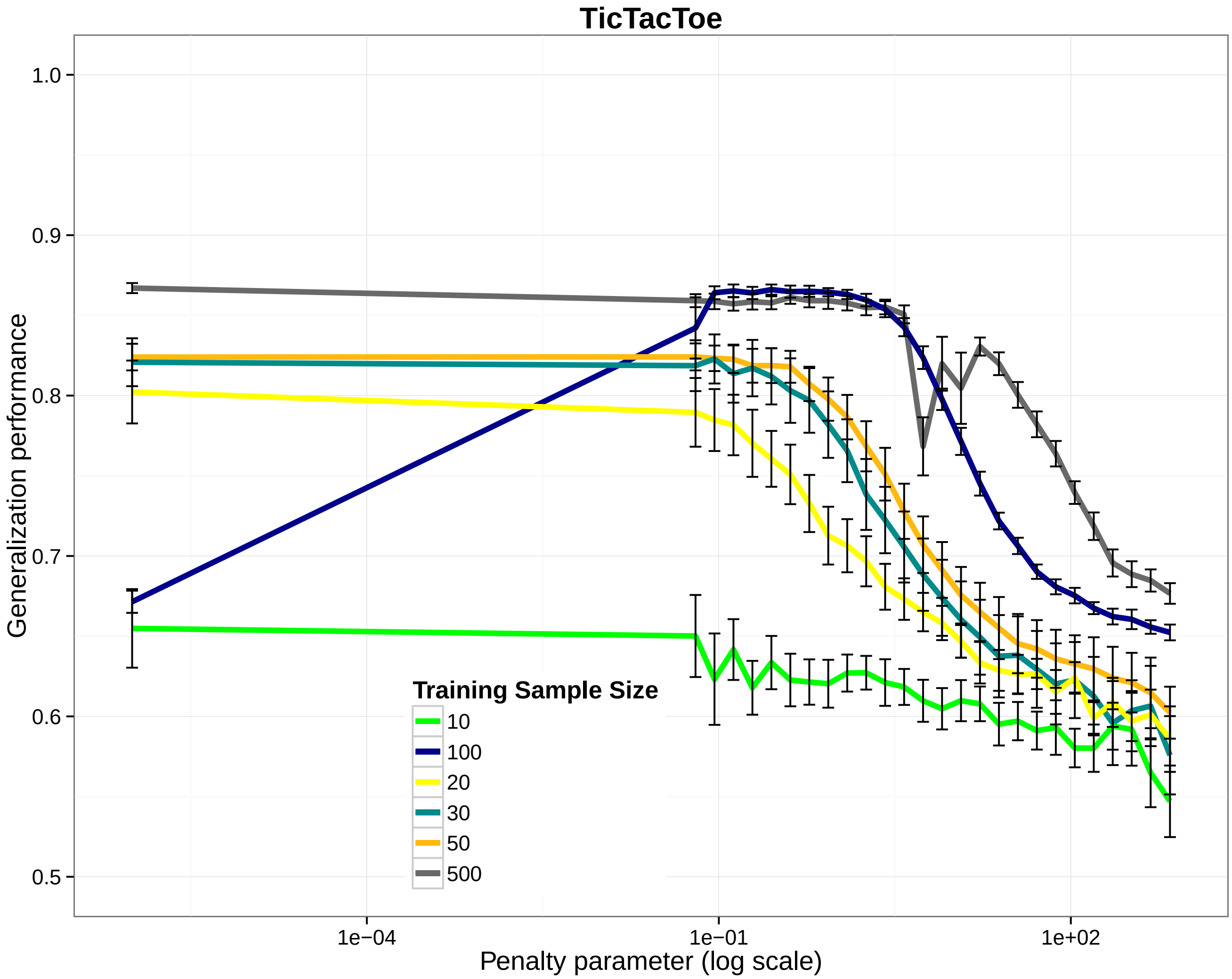

Figure 10: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, 100, and 500 of all instances.

1.11 Bupa

Number of instances: 345
Number of attributes: 6
Attribute Information:
1. Mcv 2. Alkphos 3. Sgpt 4. Sgot 5. Gammagt 6. Drinks, Class: positive,negative
This data set analizes some liver disorders that might arise from excessive alcohol consumption (the first 5 variables), and the number of half-pint equivalents of alcoholic beverages drunk per day for each individual. The task is to select if a given individual suffers from alcoholism.
Average (abs) corr: 0.26. Min (abs) corr: 0.04. Max (abs) corr: 0.74

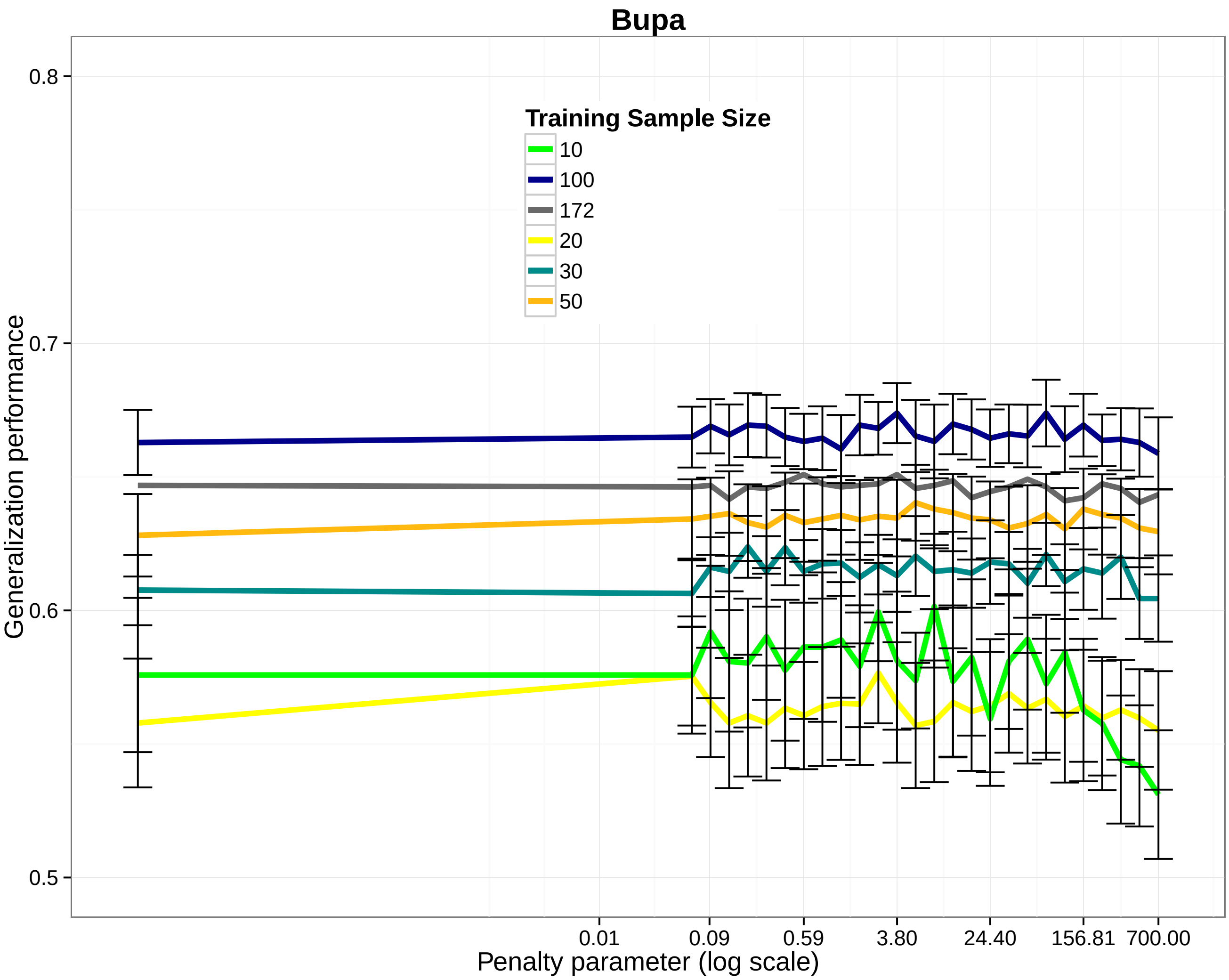

Figure 11: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, 100, and 172 (50%) of all instances.

1.12 Phoneme

Number of instances: 5404
Number of attributes: 5
Attribute Information:
1. Aa 2. Ao 3. Dcl 4. ly 5. Sh Class: 0,1
The aim of this dataset is to distinguish between nasal (class 0) and oral sounds (class 1).
Average (abs) corr: 0.13. Min (abs) corr: 0.007. Max (abs) corr: 0.32

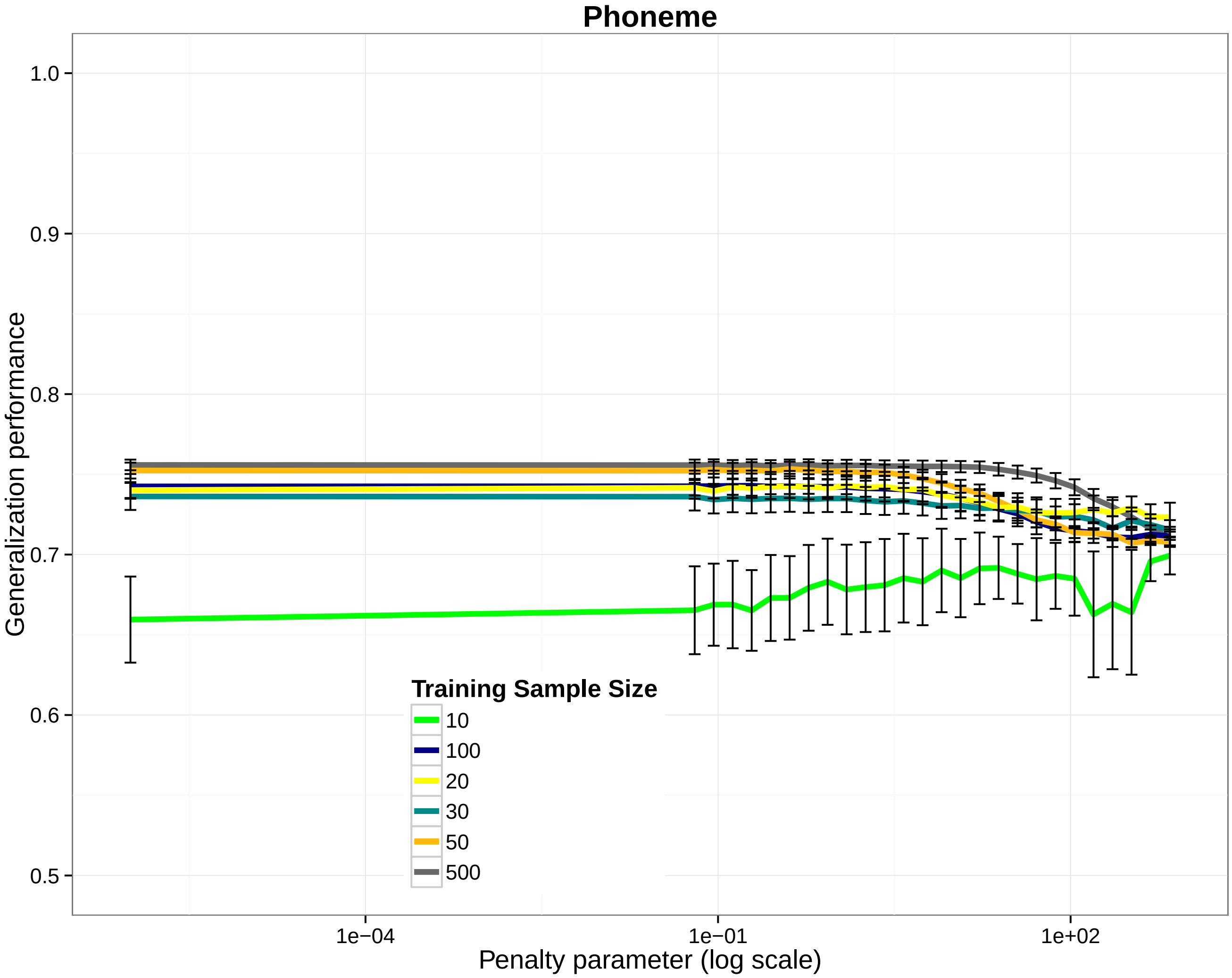

Figure 12: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, 100, and 500 of all instances.

1.13 Banana

Number of instances: 5300
Number of attributes: 2

Attribute Information:
1. At1 2. At2 Class: 1,-1
An artificial data set where instances belongs to several clusters with a banana shape. There are two attributes At1 and At2 corresponding to the x and y axis, respectively. The class label (-1 and 1) represents one of the two banana shapes in the dataset.
Average (abs) corr: 0.13. Min (abs) corr: 0.007. Max (abs) corr: 0.32

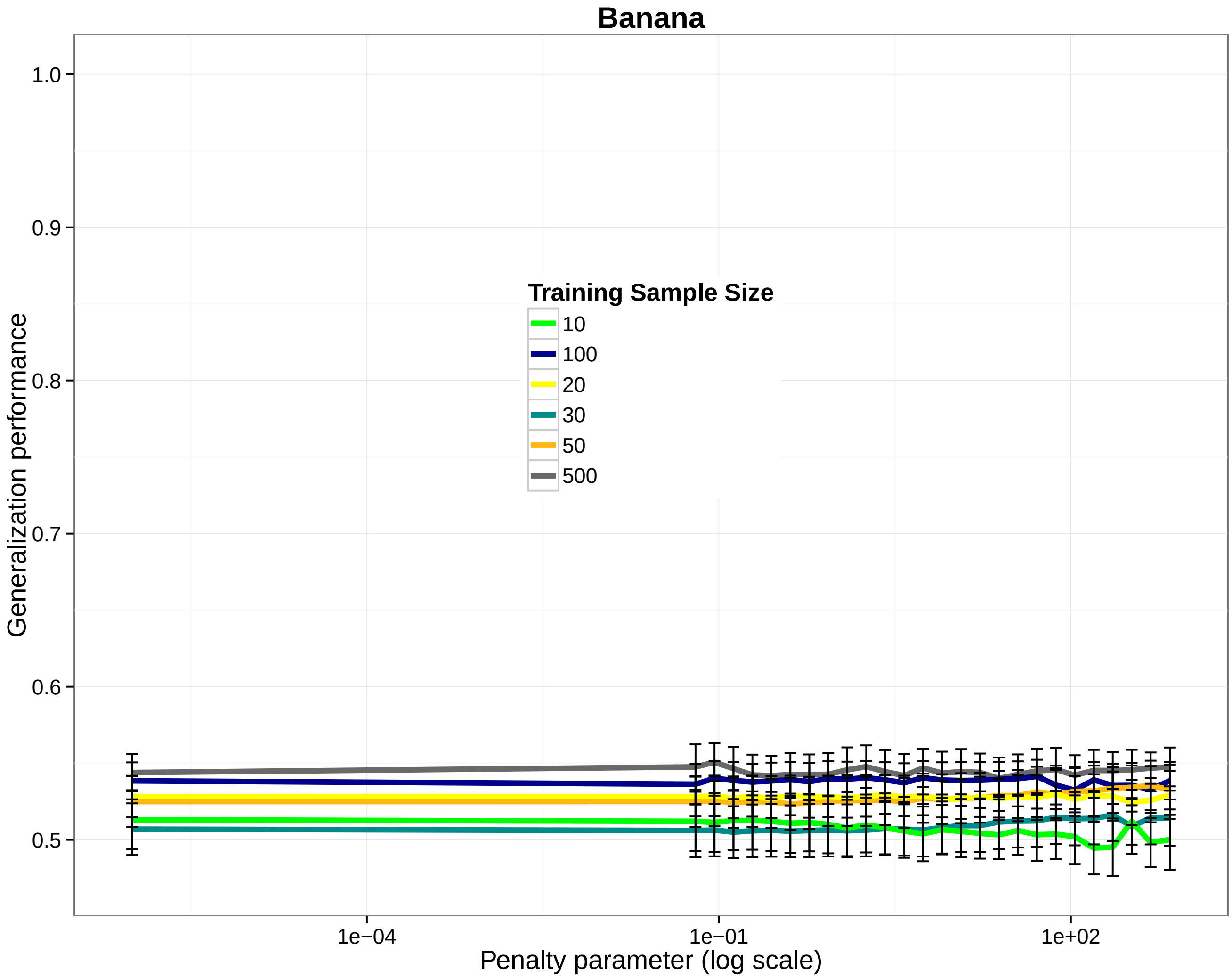

Figure 13: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, 100, and 500 of all instances.

1.14 Wholesale

Number of instances: 440
Number of attributes: 6
Attribute Information:

1) FRESH: annual spending (m.u.) on fresh products (Continuous); 2) MILK: annual spending (m.u.) on milk products (Continuous); 3) GROCERY: annual spending (m.u.)on grocery products (Continuous); 4) FROZEN: annual spending (m.u.)on frozen products (Continuous) 5) DETERGENTS PAPER: annual spending (m.u.) on detergents and paper products (Continuous) 6) DELICATESSEN: annual spending (m.u.)on and delicatessen products (Continuous); Class CHANNEL: Channel - Horeca (Hotel/Restaurant/Cafe) or Retail channel (-1, +1)

The data set refers to clients of a wholesale distributor. The annual spending in monetary units (m.u.) on diverse product categories is used to predict the channel.
Average (abs) corr: 0.30. Min (abs) corr: 0.011. Max (abs) corr: 0.93

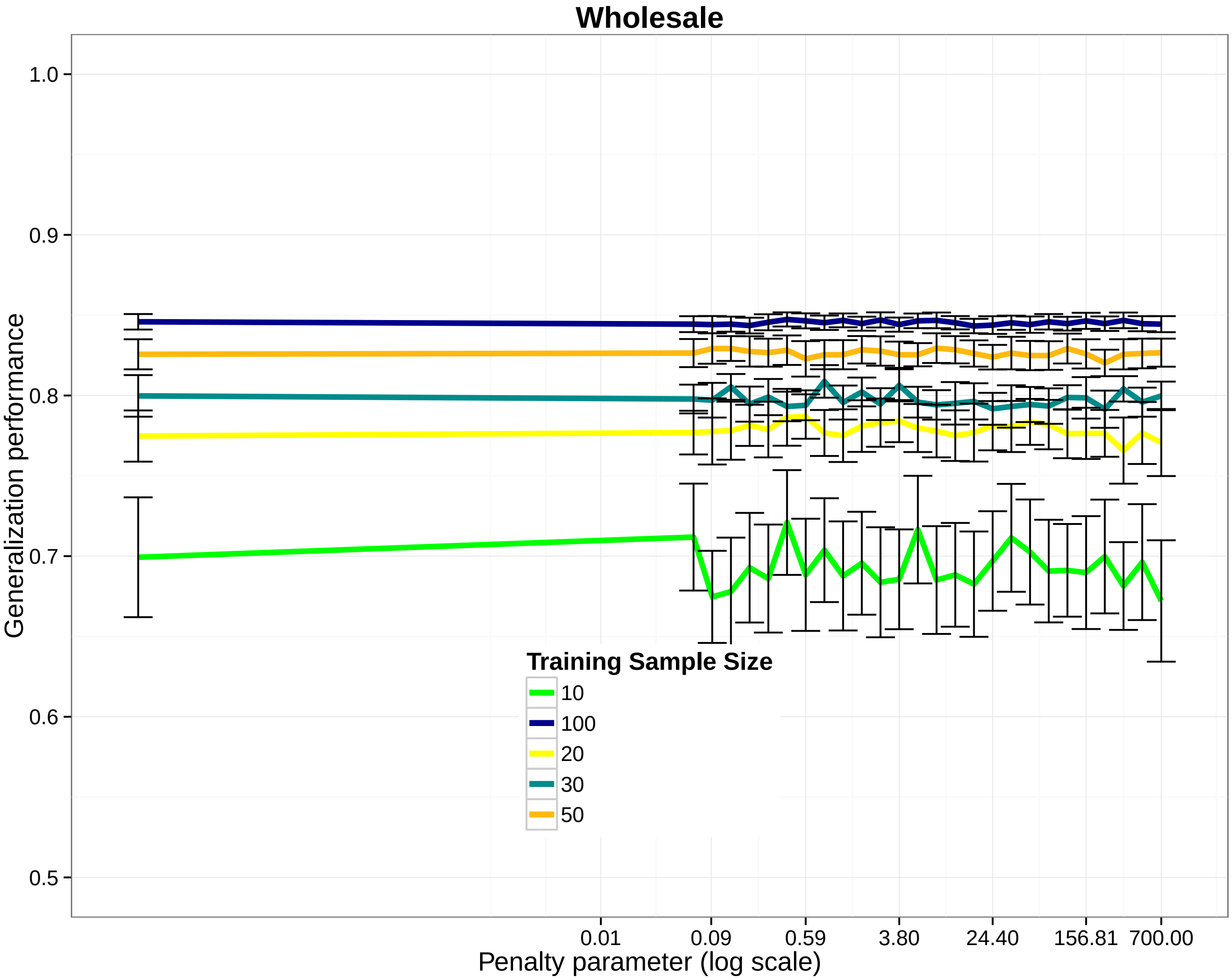

Figure 14: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, 100 of all instances.

1.15 Appendicitis

Number of instances: 106
Number of attributes: 7
Attribute Information:
At1 [0.0,1.0] At2 [0.0,1.0] At3 [0.0,1.0] At4 [0.0,1.0] At5 [0.0,1.0] At6 [0.0,1.0] At7 [0.0,1.0] Class (0,1)

The reason that a training sample of 100 does worse here is that the test set only contains 6 items in that condition.
This dataset was proposed in S. M. Weiss, and C. A. Kulikowski, Computer Systems That Learn (1991).The data represents 7 medical measures taken over 106 patients on which the class label represents if the patient has appendicitis (class label 1) or not (class label 0).
Average (abs) corr: 0.48. Min (abs) corr: 0.01. Max (abs) corr: 0.98

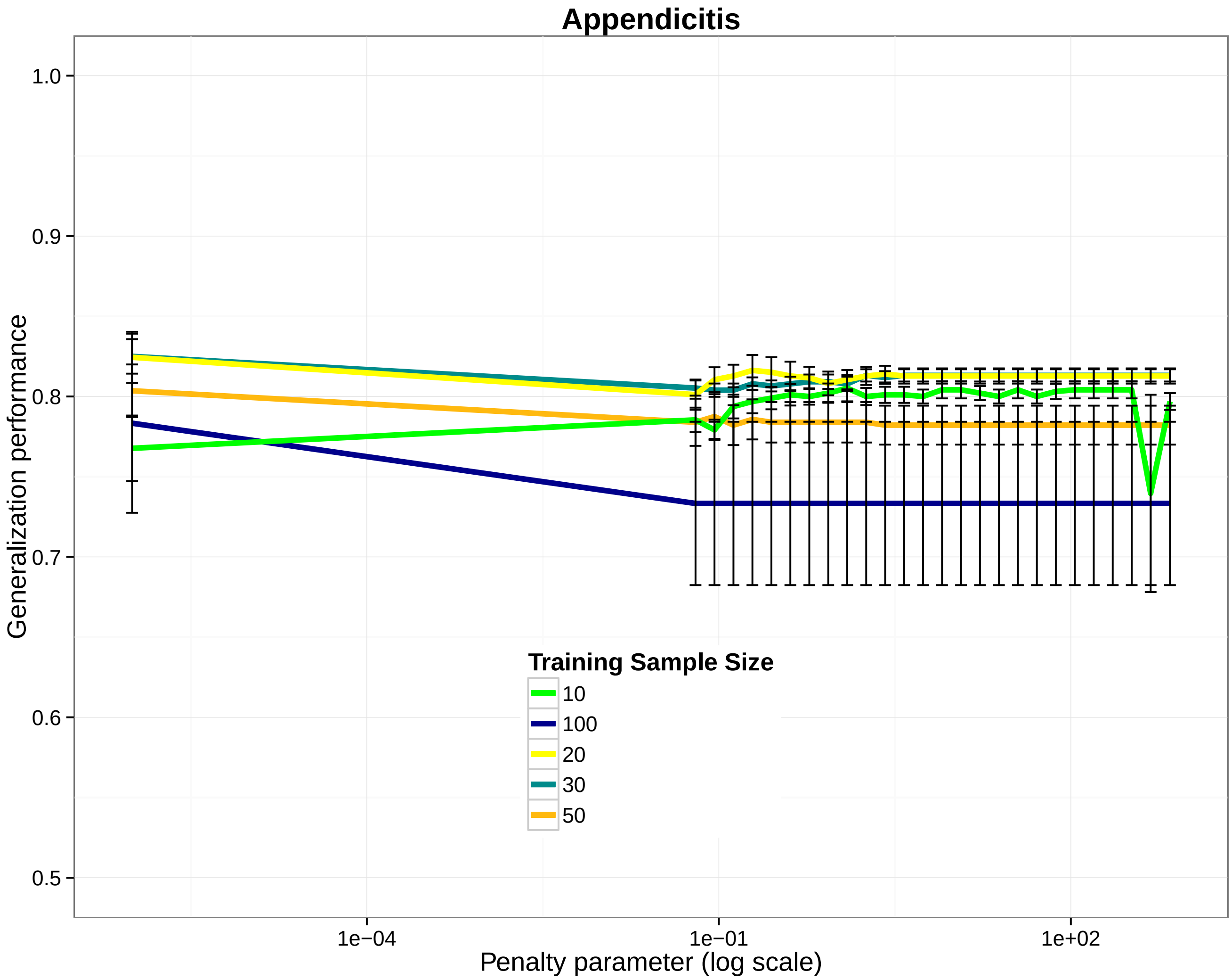

Figure 15: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50, 100 of all instances.

1.16 Blood Transfer

Number of instances: 748
Number of attributes: 4
Attribute Information:
R (Recency - months since last donation),
F (Frequency - total number of donation),
M (Monetary - total blood donated in c.c.),
T (Time - months since first donation),
Class: a binary variable representing whether he/she donated blood in March 2007 (1, 0).

Average (abs) corr: 0.47. Min (abs) corr: 0.16. Max (abs) corr: 1

(Problem: There is a complete redundancy between 2 attributes. Even at penalty = 0 the weights on the 4 attributes are very small: $[-0.004624226 \quad -0.005250064 \quad -1.14519E{-}06 \quad 0.001945579]$ and almost non-predictive. Weights are even more unstable when training samples are small.)

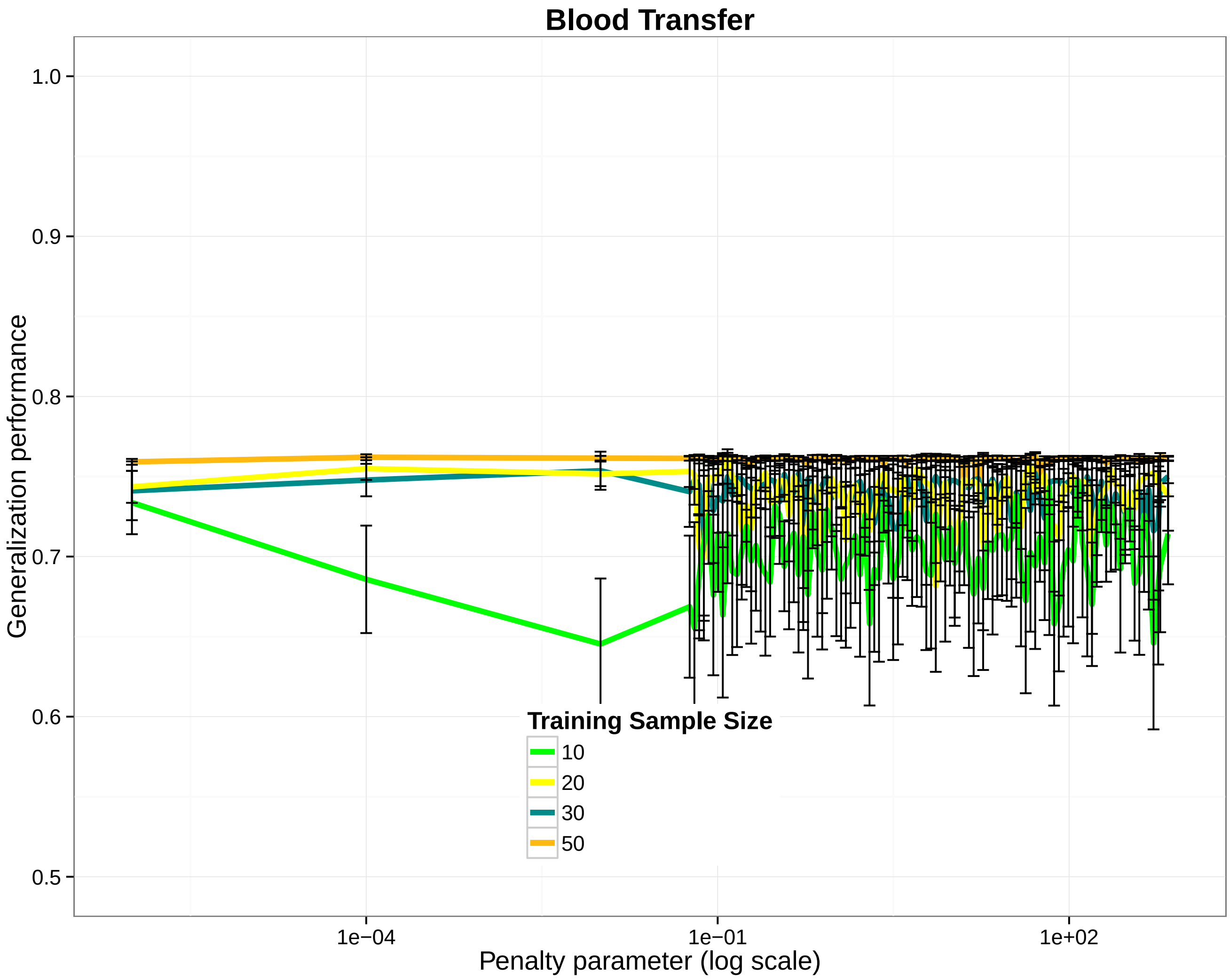

Figure 16: Performance of the COR model as a function of the penalization parameter. Training sample size varied between 10, 20, 30, 50 of all instances.