

Randomer Forests

Anonymous Author(s)

Affiliation

Address

email

Abstract

Insert abstract

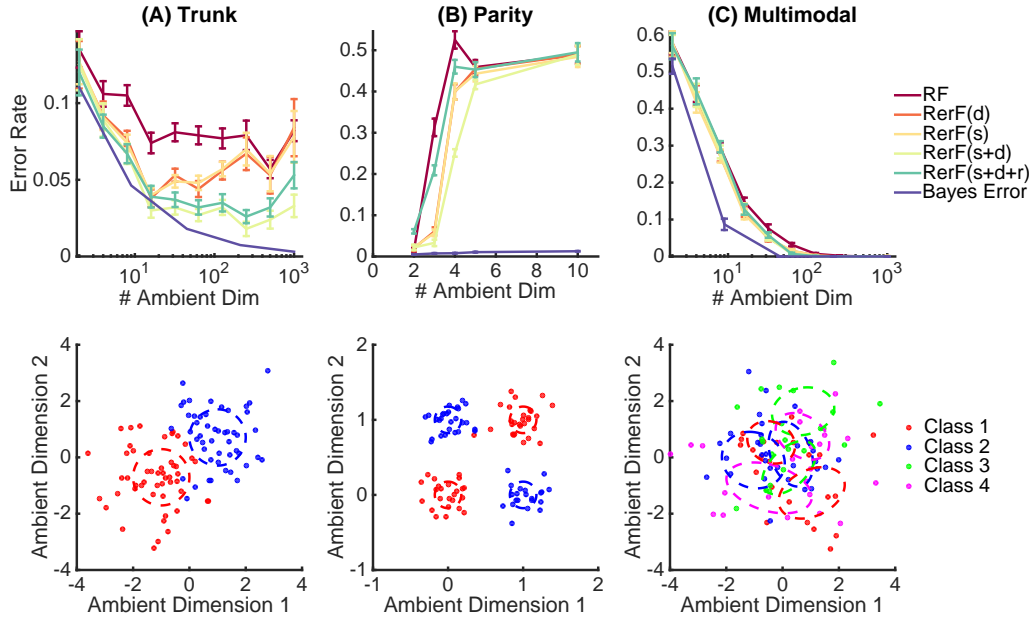


Figure 1: Classification performance comparing Random Forest (RF) to several variants of Randomer Forest (R'er F), and Bayes optimal performance, on three distinct simulation settings: (A) Trunk, (B) Parity, and (C) Multimodal (see Methods for details). For all settings, the top panel depicts misclassification rate vs. the number of ambient (coordinate) dimensions, and the bottom panel shows a 2D scatter plot of the first 2 coordinates (dashed circles denote the standard deviation level set). Note that in all settings, for all number of dimensions, R'er F outperforms RF, even Trunk and Parity, which were designed specifically for RF because the discriminant boundary naturally lies along the coordinate basis.

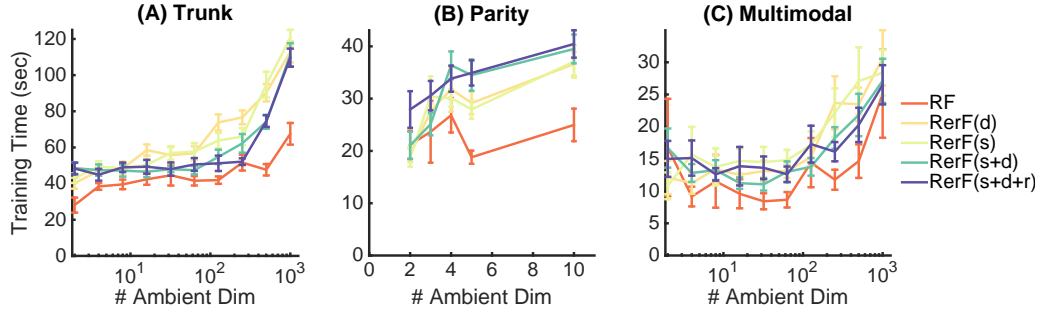


Figure 2: Classifier training time comparing RF to several variants of R'er F, same setting as the top row of Figure 1. The only difference is that the y-axis here labels training time (in seconds). Although R'er F requires slightly more time than RF (largely due to random sampling of projection matrices), they scale similarly.

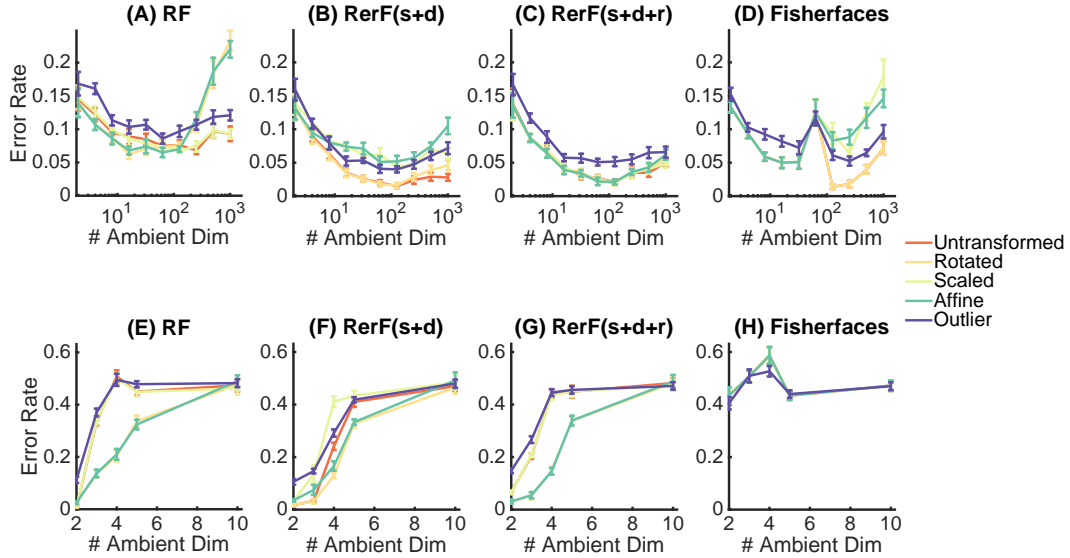


Figure 3: The effect of various transformations applied to the Trunk simulation (see Methods for details) on classification performance of (A) Random Forest, (B) Randomer Forest (sparse+delta) and (C) Randomer Forest (sparse+delta+robust). Except for scaling, classification performance of RF is compromised by all transformations. R'er F (s+d) is invariant to rotation, and R'er F (s+d+r) is invariant to rotation, scaling, and affine transformations.

108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161

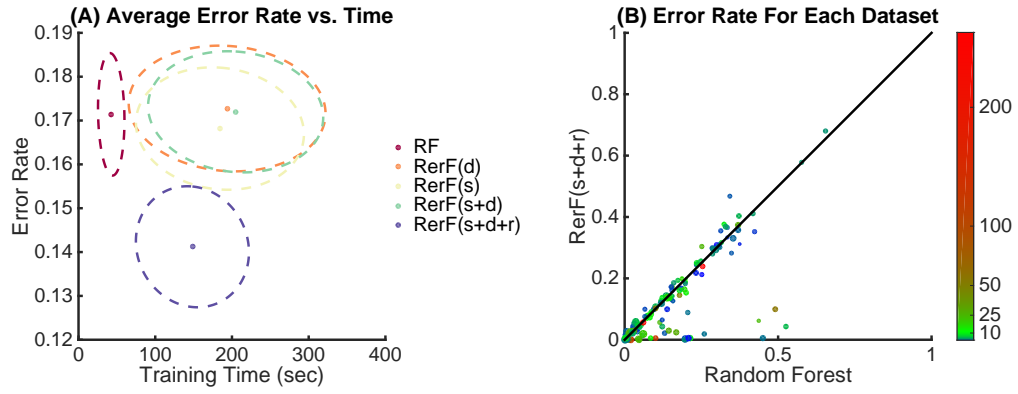


Figure 4: (A) Classification error and training time of Random Forest and Randomer Forests averaged over 121 real datasets. Dashed lines represent a 0.1 SD level curve. (B) Classification error of Randomer Forest (sparse+delta+robust) vs. that of Random Forest for each of the 121 datasets. The black line indicates equal classification error of the two algorithms. Color indicates dimensionality of the datasets and size of points indicates number of samples.