

---

# Orthogonal Gradient Boosting for Interpretable Additive Rule Ensembles

## Supplementary Information

---

Anonymous Author(s)

Affiliation

Address

email

## 1 A Proofs

### 2 A.1 Proof of Lemma 4.1

3 *Proof.* Let  $\mathbf{f} = [\mathbf{Q}; \mathbf{g}] \alpha$  and  $\tilde{\mathbf{f}} = [\mathbf{Q}; \mathbf{q}] \beta$ . We can decompose the squared error

$$\begin{aligned} \|\mathbf{f} - \tilde{\mathbf{f}}\|^2 &= \|[\mathbf{Q}; \mathbf{g}] \alpha - [\mathbf{Q}; \mathbf{q}] \beta\|^2 \\ &= \|[\mathbf{Q}; \mathbf{g}_{\parallel} + \mathbf{g}_{\perp}] \alpha - [\mathbf{Q}; \mathbf{q}_{\parallel} + \mathbf{q}_{\perp}] \beta\|^2 \\ &= \|[\mathbf{Q}; \mathbf{g}_{\parallel}] \alpha + \alpha_t \mathbf{g}_{\perp} - [\mathbf{Q}; \mathbf{q}_{\parallel}] \beta + \beta_t \mathbf{q}_{\perp}\|^2 \\ &= \|[\mathbf{Q}; \mathbf{g}_{\parallel}] \alpha - [\mathbf{Q}; \mathbf{q}_{\parallel}] \beta\|^2 + \|\alpha_t \mathbf{g}_{\perp} - \beta_t \mathbf{q}_{\perp}\|^2 \end{aligned}$$

4 where the last step follows from the Pythagorean theorem and the fact that  $\alpha_t \mathbf{g}_{\perp} - \beta_t \mathbf{q}_{\perp}$  is an element  
5 from the orthogonal complement of  $\text{range}[\mathbf{Q}; \mathbf{g}_{\parallel}] = \text{range}[\mathbf{Q}; \mathbf{g}_{\parallel}] = \text{range } \mathbf{Q}$ . The equality of these  
6 ranges also implies that  $\beta_1, \dots, \beta_{t-1}$  can, for all choices of  $\beta_t$ , be chosen such that the left term of  
7 the error decomposition is 0. Setting  $\gamma = \beta_t / \alpha_t$ , it follows that

$$\begin{aligned} \min_{\beta \in \mathbb{R}^t} \|\mathbf{f} - \tilde{\mathbf{f}}\|^2 &= \min_{\beta \in \mathbb{R}^t} \|\alpha_t \mathbf{g}_{\perp} - \beta_t \mathbf{h}_{\perp}\|^2 \\ &= \min_{\gamma \in \mathbb{R}^t} \alpha_t^2 \|\mathbf{g}_{\perp} - \gamma \mathbf{q}_{\perp}\|^2 \\ &= \min_{\gamma \in \mathbb{R}^t} \alpha_t^2 (\|\mathbf{g}_{\perp}\|^2 - 2\gamma \mathbf{q}_{\perp}^T \mathbf{g}_{\perp} + \gamma^2 \|\mathbf{q}_{\perp}\|^2) \end{aligned}$$

8 and plugging in the minimizing  $\gamma = \mathbf{q}_{\perp}^T \mathbf{g}_{\perp} / \|\mathbf{q}_{\perp}\|^2$

$$= \alpha^2 (\|\mathbf{g}_{\perp}\|^2 - (\mathbf{g}_{\perp}^T \mathbf{q}_{\perp})^2 / \|\mathbf{q}_{\perp}\|^2) ,$$

9 from which, noting that  $\mathbf{g}_{\perp}^T \mathbf{q}_{\perp} = \mathbf{g}_{\perp}^T \mathbf{q}$ , the claim follows. □

### 10 A.2 Proof of Lemma 4.2

11 *Proof.* After the weight correction step  $\beta$  is a stationary point of  $R(\mathbf{Q}(\cdot))$ , i.e., we have for all  $j \in [t]$

$$0 = \frac{\partial R(\mathbf{Q}\beta)}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial l(\tilde{\mathbf{q}}_i^T \beta, y_i)}{\partial \beta_j} = \sum_{i=1}^n q_{ij} \underbrace{\frac{\partial l(\tilde{\mathbf{q}}_i^T \beta, y_i)}{\partial \tilde{\mathbf{q}}_i^T \beta}}_{g_i} = \mathbf{q}_j^T \mathbf{g} .$$

12 □

### 13 A.3 Proof of Theorem 4.3

14 The condition of Theorem 4.3 states that:

15 Let  $\mathbf{Q} \in \mathbb{R}^{n \times (t-1)}$  be the selected query matrix,  $\mathbf{g}$  the corresponding gradient vector after a full  
16 weight correction, and  $\mathbf{q}^*$  be a maximizer of the **orthogonal gradient boosting objective** function  
17 defined by

$$\text{obj}_{\text{ogb}}(q) = \frac{|\mathbf{g}^T \mathbf{q}|}{\|\mathbf{q}_{\perp}\| + \epsilon}$$

18 where  $\mathbf{q}_{\perp}$  is the projection of  $q$  onto the orthogonal complement of range  $\mathbf{Q}$ .

#### 19 A.3.1 Property a

20 **Proposition A.1.** For  $\epsilon \rightarrow 0$ ,  $[\mathbf{q}_1, \dots, \mathbf{q}_{t-1}, \mathbf{q}^*]$  is the best approximation to  $[\mathbf{q}_1, \dots, \mathbf{q}_{t-1}, \mathbf{g}]$ .

*Proof.* If  $\epsilon \rightarrow 0$ , then  $\text{obj}_{\text{ogb}}(q) \rightarrow \frac{|g^T q|}{\|q_{\perp}\|}$ . If  $\mathbf{q}^*$  is a maximizer of  $\text{obj}_{\text{ogb}}$ , then as shown in Lemma  
4.1,  $\mathbf{q}^*$  minimises the minimum distance from all

$$\mathbf{f} \in \text{span}\{\mathbf{q}_1, \dots, \mathbf{q}_{t-1}, \mathbf{g}\}$$

to the subspace of

$$\text{span}\{\mathbf{q}_1, \dots, \mathbf{q}_{t-1}, \mathbf{q}^*\}.$$

21 Therefore, the subspace spanned by  $[\mathbf{q}_1, \dots, \mathbf{q}_{t-1}, \mathbf{q}^*]$  is the best approximation to the subspace  
22 spanned by  $[\mathbf{q}_1, \dots, \mathbf{q}_{t-1}, \mathbf{g}]$ .  $\square$

#### 23 A.3.2 Property b

24 **Proposition A.2.** For  $\epsilon \rightarrow \infty$ ,  $\mathbf{q}^*$  is also a maximizer of  $\text{obj}_{\text{gs}}$  and any maximizer of  $\text{obj}_{\text{gs}}$  is also a  
25 maximizer of  $\text{obj}_{\text{ogb}}$ .

26 *Proof.* Let  $q_1$  and  $q_2$  be any two queries and denote by  $\text{obj}_{\text{ogb}}^{(\epsilon)}(q)$  the  $\text{obj}_{\text{ogb}}$ -value of  $q$  for a specific  
27  $\epsilon$ . Then

$$\begin{aligned} & \lim_{\epsilon \rightarrow \infty} \epsilon \left( \text{obj}_{\text{ogb}}^{(\epsilon)}(q_1) - \text{obj}_{\text{ogb}}^{(\epsilon)}(q_2) \right) \\ &= \lim_{\epsilon \rightarrow \infty} \epsilon \left( \frac{|g^T \mathbf{q}_1|}{\|\mathbf{q}_1^{\perp}\| + \epsilon} - \frac{|g^T \mathbf{q}_2|}{\|\mathbf{q}_2^{\perp}\| + \epsilon} \right) \\ &= \lim_{\epsilon \rightarrow \infty} \left( \frac{|g^T \mathbf{q}_1|}{\|\mathbf{q}_1^{\perp}\|/\epsilon + 1} - \frac{|g^T \mathbf{q}_2|}{\|\mathbf{q}_2^{\perp}\|/\epsilon + 1} \right) \\ &= |g^T \mathbf{q}_1| - |g^T \mathbf{q}_2| \\ &= \text{obj}_{\text{gs}}(q_1) - \text{obj}_{\text{gs}}(q_2) \end{aligned}$$

28 Thus for large enough  $\epsilon$ , the signs of  $\text{obj}_{\text{ogb}}^{(\epsilon)}(q_1) - \text{obj}_{\text{ogb}}^{(\epsilon)}(q_2)$  and  $\text{obj}_{\text{gs}}(q_1) - \text{obj}_{\text{gs}}(q_2)$  agree.  
29 Therefore, a query  $q$  is a  $\text{obj}_{\text{gs}}$ -maximizer, i.e.,  $\text{obj}_{\text{gs}}(q) \geq \text{obj}_{\text{gs}}(q')$  for all  $q' \in \mathcal{Q}$ , if and only if  $q$   
30 is a  $\text{obj}_{\text{ogb}}$ -maximizer, i.e.,  $\text{obj}_{\text{ogb}}(q) \geq \text{obj}_{\text{ogb}}(q')$  for all  $q' \in \mathcal{Q}$ .  $\square$

#### 31 A.3.3 Property c

32 **Proposition A.3.** For  $\epsilon = 0$  and  $\|\mathbf{q}_{\perp}\| > 0$ , the ratio  $(\frac{\text{obj}_{\text{ogb}}(q)}{\text{obj}_{\text{gb}}(q)})^2$  is equal to  $1 + (\frac{\|\mathbf{q}_{\parallel}\|}{\|\mathbf{q}_{\perp}\|})^2$ .

33 *Proof.* If  $\epsilon = 0$  and  $\|\mathbf{q}_\perp\| > 0$ , then

$$\begin{aligned} \left( \frac{\text{obj}_{\text{ogb}}(q)}{\text{obj}_{\text{gb}}(q)} \right)^2 &= \frac{\frac{|\mathbf{g}^T \mathbf{q}|^2}{\|\mathbf{q}_\perp\|^2}}{\frac{|\mathbf{g}^T \mathbf{q}|^2}{\|\mathbf{q}\|^2}} = \frac{\|\mathbf{q}\|^2}{\|\mathbf{q}_\perp\|^2} \\ &= \frac{\|\mathbf{q}_\parallel\|^2 + \|\mathbf{q}_\perp\|^2}{\|\mathbf{q}_\perp\|^2} \\ &= 1 + \left( \frac{\|\mathbf{q}_\parallel\|}{\|\mathbf{q}_\perp\|} \right)^2 \end{aligned}$$

34

□

#### 35 **A.3.4 Property d**

36 **Proposition A.4.** The objective value  $\text{obj}_{\text{ogb}}(q)$  is upper bounded by  $\|\mathbf{g}\|$ .

37 *Proof.* If we divide the numerator and denominator of  $\text{obj}_{\text{ogb}}(\mathbf{q})$  with  $\|\mathbf{q}_\perp\|$ , then we can get

$$\begin{aligned} \text{obj}_{\text{ogb}}(\mathbf{q}) &= \frac{|\mathbf{g}^T \mathbf{q}|}{\|\mathbf{q}_\perp\| + \epsilon} \\ &= \frac{\frac{|\mathbf{g}^T \mathbf{q}_\perp|}{\|\mathbf{q}_\perp\|}}{1 + \frac{\epsilon}{\|\mathbf{q}_\perp\|}} \end{aligned}$$

38 according to the Cauchy–Schwarz inequality,  $\frac{|\mathbf{g}^T \mathbf{q}|}{\|\mathbf{q}_\perp\|} \leq \frac{\|\mathbf{g}\| \|\mathbf{q}_\perp\|}{\|\mathbf{q}_\perp\|} = \|\mathbf{g}\|$ , so,

$$\text{obj}_{\text{ogb}}(\mathbf{q}) \leq \frac{\|\mathbf{g}\|}{1 + \frac{\epsilon}{\|\mathbf{q}_\perp\|}}$$

39 as  $\|\mathbf{q}_\perp\|$  is upper bounded by the number of data points  $n$ ,

$$\begin{aligned} \text{obj}_{\text{ogb}}(\mathbf{q}) &\leq \frac{\|\mathbf{g}\|}{1 + \frac{\epsilon}{n}} \\ \text{obj}_{\text{ogb}}(\mathbf{q}) &\leq \|\mathbf{g}\|. \end{aligned}$$

40

□

#### 41 **A.4 Proof of Theorem 4.4**

42 *Proof.* To see the claim, we first rewrite the objective value for the  $i$ -th prefix as

$$\frac{\mathbf{g}^T \mathbf{q}^{(i)}}{\|\mathbf{q}_\perp^{(i)}\| + \epsilon} = \frac{\mathbf{g}^T \mathbf{q}^{(i)}}{\|\mathbf{q}^{(i)}\| - \|\mathbf{q}_\parallel^{(i)}\| + \epsilon}.$$

43 The value of  $\|\mathbf{q}^{(i)}\|$  is trivially given as  $\sqrt{i}$ , and  $\mathbf{g}^T \mathbf{q}^{(i)}$  can be easily computed for all  $i \in [L]$  in time  
44  $O(n)$  via cumulative summation. Finally we can reduce the problem of computing the (squared)  
45 norms of the  $l$  projected prefixes to computing the  $t$  (squared) norms of the prefixes on the subspaces  
46 given by the individual orthonormal basis vectors via

$$\|\mathbf{q}_\parallel^{(i)}\|^2 = \left\| \sum_{k=1}^t \mathbf{o}_k \mathbf{o}_k^T \mathbf{q}^{(i)} \right\|^2 = \sum_{k=1}^t \|\mathbf{o}_k \mathbf{o}_k^T \mathbf{q}^{(i)}\|^2.$$

Each of these  $t$  sequences of (squared) norms can be computed in time  $O(n)$  by rewriting

$$\begin{aligned}\|\mathbf{o}_k \mathbf{o}_k^T \mathbf{q}^{(i)}\| &= \left\| \mathbf{o}_k \mathbf{o}_k^T \left( \sum_{j=1}^i \mathbf{e}_{\sigma(j)} \right) \right\| \\ &= \|\mathbf{o}_k\| \sum_{j=1}^i \mathbf{o}_k^T \mathbf{e}_{\sigma(j)} \\ &= \sum_{j=1}^i o_{k, \sigma(j)}\end{aligned}$$

where the last equality shows how an  $O(n)$ -computation is achieved via cumulative summation of the  $k$ -th basis vector elements in the order given by  $\sigma$ .  $\square$

## B Greedy approximation to bounding function

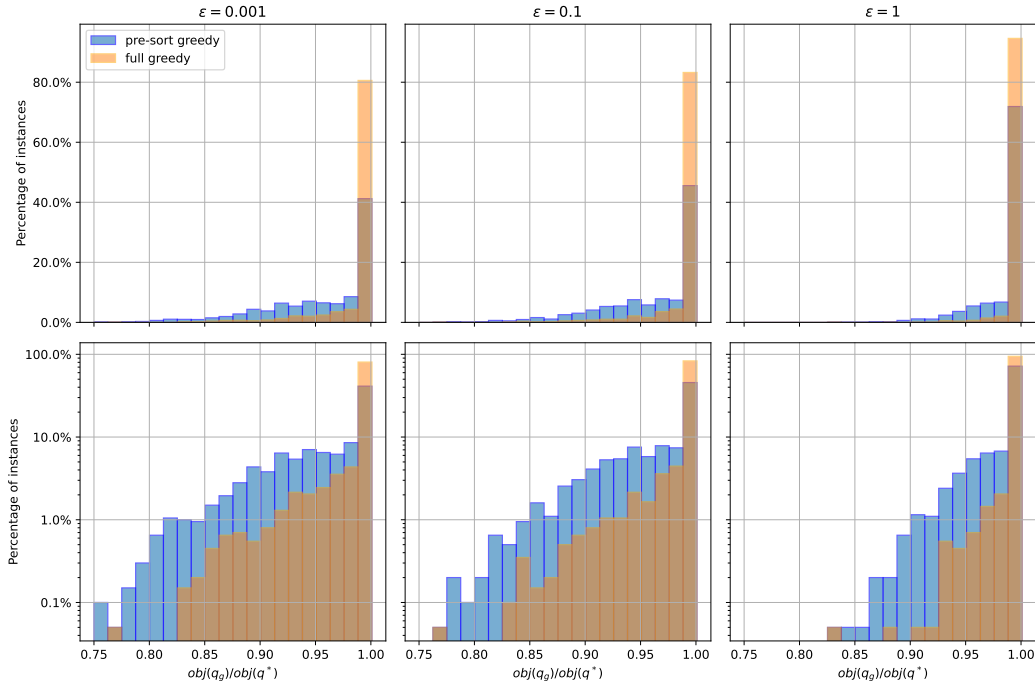


Figure 4: The number of instances of ratios between the best objective values obtained from the greedy search and the true optimal objective value. The upper figures are in linear scales and the lower figures are in log scales. The total variation distances for these three values of  $\epsilon$  are 0.394, 0.377 and 0.227.

The branch-and-bound search described in Section 3.3 requires an efficient way of calculating the value of  $\text{bnd}(\mathbf{q}) = \max\{\text{obj}(\mathbf{q}') : \mathbf{q}' \leq \mathbf{q}, \mathbf{q}' \in \{0, 1\}^n\}$ . It is too expensive to enumerate all possible  $\mathbf{q}'$ s as there are  $2^n$  cases in the worst case. One solution to this problem is that we can relax the constraint  $\mathbf{q}' \in \{0, 1\}^n$  to  $\mathbf{q}' \in [0, 1]^n$  and it can be solved by quadratic programming. However, this relaxation is too loose and inefficient. Instead, we consider relaxing the admission constraint and solve the problem using greedy algorithms.

A full greedy approach can be used to approximate the maximum objective value of the subset of data points selected by  $\mathbf{q}$ , which is the bounding value  $\text{bnd}(\mathbf{q})$ . Given a query  $\mathbf{q}^{(t-1)} \leq \mathbf{q}$ , we need to find the data point selected by  $\mathbf{q}$  which maximise the objective function, and use it with  $\mathbf{q}^{(t-1)}$  to

60 form a  $\mathbf{q}'(t)$ .

$$i_*^{(t)} = \arg \max_{i \in I(\mathbf{q}) - I(\mathbf{q}'^{(t-1)})} \frac{\mathbf{g}^T(\mathbf{q}'^{(t-1)} + \mathbf{e}_i)}{\|(\mathbf{q}'^{(t-1)} + \mathbf{e}_i)_\perp\| + \epsilon}.$$

61 where  $I(\mathbf{q}) = \{i : \mathbf{q}(x_i) = 1, 1 \leq i \leq n\}$ ,  $0 \leq t \leq |I(\mathbf{q})|$ ,  $\mathbf{q}'^{(0)} = \mathbf{0}$  and  $\mathbf{q}'^{(t)} = \mathbf{q}'^{(t-1)} + \mathbf{e}_{i_*^{(t)}}$ .

62 We use the maximum value of  $\text{obj}(\mathbf{q}'^{(t)})$  as the bounding value for query  $\mathbf{q}$ . The computation time  
63 complexity level of this approach is  $O(n^2)$  for each query, which is not as efficient as the presorting  
64 greedy approach described in Section 4.3.

65 The presorting greedy approach of solving the prefix optimization problem described in Section 4.3  
66 leads to another approximation to the optimal objective function value for the queries which cover  
67 subsets of data points covered by  $\mathbf{q}$ . As proved in Theorem 4.4, this approach has a time complexity  
68 of  $O(tn)$ .

69 We test 2000 instances for different initial queries and initial gradient values to see the difference  
70 between the approximation of  $\text{bnd}(\mathbf{q})$  obtained by the full greedy approach, the pre-sorting greedy  
71 approach, and the actual optimal objective values (obtained by a brute-force approach). We choose  
72 three different values of  $\epsilon$ : 0.001, 0.1 and 1.

73 Figure 4 compares the ratio between the approximations to  $\text{bnd}(\mathbf{q})$  obtained by the two greedy  
74 approaches and the true optimal objective value. The Y axis of Figure 4 represents the percentage of  
75 instances of different ratios.

76 According to the comparison, the full greedy approach can approximate the true bounding function  
77 better than the presorting greedy approach. For smaller  $\epsilon$  values ( $\epsilon = 0.001$ ), there are 90% instances  
78 whose approximation values are more than 90% of the true bounding function values, while 96% of  
79 instances approximate more than 90% of the value of  $\text{bnd}(\mathbf{q})$  using the full greedy approach. For  
80  $\epsilon = 0.1$ , both algorithms have slight better (both 1% promotion) approximation than  $\epsilon = 0.001$ .  
81 It can be observed that for  $\epsilon = 1$ , both algorithms have more instances where the approximations  
82 are closed to the true bounding values. However, if the value of  $\epsilon$  is too large, then the calculated  
83 objective values cannot be accurate according to Theorem 4.3. Comparing the statistical distances of  
84 these two greedy approaches, it is reasonable to use the presorting greedy approach to approximate  
85 the bounding values.

86 To approximate the true bounding function more efficiently and more accurate, we adopt the presorting  
87 greedy approach in this research.

## 88 C Experiments configurations

89 The experiments are conducted on a computer with CPU 'Intel(R) Core(TM) i5-10300H CPU @  
90 2.50GHz' and memory of 24G.

## 91 D Comparison of scores

92 Table 2 shows the comparison of the area under the score / cognitive complexity curve of SIRUS,  
93 Gradient Sum, Gradient boosting, XGBoost and FCOGB for benchmark datasets of classification,  
94 regression and Poisson regression problems. All the experiments have the same configurations as  
95 discussed in Section 5.

Table 2: Comparison of the area under the score / cognitive complexity curve of SIRUS, Gradient Sum(GS), Gradient boosting (GB), XGBoost (XGB) and FCOGB for benchmark datasets of classification (upper), regression (middle) and Poisson regression problems (lower).

Dataset	$d$	$n$	Training Scores / CC AUC					Testing Scores / CC AUC				
			SIRUS	GS	GB	XGB	FCOGB	SIRUS	GS	GB	XGB	FCOGB
titanic	7	1043	.783	.858	.846	.856	<b>.866</b>	.803	<b>.826</b>	.805	.800	.816
tic-tac-toe	27	958	.677	.783	.878	.888	<b>.897</b>	.669	.729	.822	.856	<b>.867</b>
iris	4	150	.854	.962	.960	.952	<b>.964</b>	.781	<b>.941</b>	.909	.925	.913
breast	30	569	.930	<b>.965</b>	.961	.964	.960	.890	<b>.964</b>	.935	.946	.951
wine	13	178	.929	<b>.975</b>	.954	.966	.974	.912	.962	.905	.931	<b>.976</b>
ibm hr	32	1470	.700	.426	.628	.690	<b>.731</b>	<b>.696</b>	.417	.591	.634	.679
telco churn	18	7043	.756	.796	.790	.787	<b>.798</b>	.776	<b>.808</b>	.794	.795	.798
gender	20	3168	<b>.925</b>	.243	.288	.288	.202	<b>.926</b>	.240	.285	.285	.200
banknote	4	1372	.874	.968	.965	<b>.969</b>	.967	.870	.967	.965	.967	<b>.967</b>
liver	6	345	.701	.812	.792	.807	<b>.812</b>	<b>.699</b>	.687	.650	.622	.576
magic	10	19020	.602	.778	.805	.806	<b>.806</b>	.605	.788	.812	<b>.814</b>	.813
adult	11	30162	.726	.731	.838	.842	<b>.847</b>	.722	.725	.828	.832	<b>.837</b>
digits5	64	3915	.423	.465	.732	.741	<b>.828</b>	.360	.451	.715	.728	<b>.797</b>
insurance	6	1338	.706	.682	.794	.795	<b>.824</b>	.633	.683	.782	.770	<b>.791</b>
friedman1	10	2000	<b>.637</b>	.458	.556	.579	.605	<b>.633</b>	.409	.431	.430	.458
friedman2	4	10000	.523	.721	.867	.873	<b>.894</b>	.536	.721	.868	.873	<b>.893</b>
friedman3	4	5000	.590	.539	.616	.630	<b>.657</b>	.547	.528	.563	.552	<b>.571</b>
wage	5	1379	.217	.308	.334	.358	<b>.383</b>	.182	<b>.244</b>	.207	.200	.180
demographics	13	6876	<b>.367</b>	.103	.168	.187	.203	<b>.389</b>	.102	.160	.174	.187
gdp	1	35	<b>.663</b>	.469	.444	.480	.513	<b>.616</b>	.434	.403	.437	.461
used cars	4	1770	.679	.766	.820	.823	<b>.848</b>	.654	.753	.810	.810	<b>.834</b>
diabetes	10	442	.448	.386	.478	.491	<b>.520</b>	<b>.377</b>	.324	.348	.318	.270
boston	13	506	.642	.493	.595	.616	<b>.642</b>	<b>.590</b>	.441	.425	.430	.426
happiness	8	315	.077	.789	.804	.802	<b>.838</b>	.032	<b>.675</b>	.541	.524	.572
life expect.	21	1649	.662	.622	.671	.685	<b>.717</b>	<b>.641</b>	.558	.603	.606	.613
mobile prices	20	2000	.696	.618	.720	.723	<b>.726</b>	<b>.711</b>	.596	.684	.671	.676
suicide rate	5	27820	<b>.287</b>	.125	.147	.153	.203	<b>.289</b>	.127	.149	.154	.204
videogame	6	16327	<b>.015</b>	.001	.001	.001	.001	<b>.016</b>	.001	.001	.001	.001
red wine	11	1599	.251	.242	.285	.304	<b>.317</b>	.177	.248	.243	.237	<b>.262</b>
covid vic	4	85	NA	.678	<b>.845</b>	NA	.827	NA	.691	<b>.865</b>	NA	.835
covid	2	225	NA	.412	.476	NA	<b>.514</b>	NA	.203	<b>.205</b>	NA	.200
bicycle	4	122	NA	.707	.778	.789	<b>.810</b>	NA	.623	.596	<b>.630</b>	.591
ships	4	34	NA	.866	.925	NA	<b>.946</b>	NA	.484	<b>.514</b>	NA	NA
smoking	2	36	NA	.677	.943	.829	<b>.957</b>	NA	.607	.756	.591	<b>.801</b>