Supplementary materials

Zhenwei Yang

Data Generating Mechanism of Example 1

Table S1: Summary of variables in example 1

Variable	**Meaning**	**Simulation**	**Categories**
P	observed predictor	$Pr(X = p) = 1/6$ $p \in \{1, 2, ..., 6\}$	6 categories, ranging from 1 to 6
U	unobserved predictor	$Pr(X = u) = 1/3$ $u \in \{1, 2, 3\}$	3 categories, ranging from 1 to 3
Y	response variable	$\tilde{Y} = 1 + 2 * P + 0.2 * U + \epsilon$ Y was formed by categorizing $\tilde{Y}$ at the boundary of Q_0, Q_1, Q_2, Q_3 and Q_4 *	4 categories, ranging from 1 to 4
ϵ	error term	$\epsilon \sim N(0, 1)$	-

* Q_0, Q_1, Q_2, Q_3 and Q_4 refer to the minimum, first quantile, second quantile, third quantile and maximum value of the data

Application on Example 3

In this section, we provide example 3 for the application of LBA-NN. The German Suicide dataset is extracted from the previous paper (Van der Heijden, Mooijaart, and De Leeuw 1992). A summary of the data is provided in Table S2. Two explanatory variables are involved in this example, gender with 2 levels and age with 17 levels. The response variable, causes of death contains 9 categories. The original dataset was split to a training set and a test set, with 43211 and 10000 observations, respectively. Three latent budgets are assigned in LBA. The optimal hyperparameters includes 32 hidden neurons, the ReLU activation for the hidden layer and the softmax activation for the output layer, learning rate of 10^{-4}.

Table S2: The contingency table for the example data 3

	Causes of Death*									
	asolliq	bgas	cothergas	dhss	edrown	fgunexp	gknives	hjump	iother	Sum
Male										
10-15	4	0	0	247	1	17	1	6	9	285
15-20	348	7	67	578	22	179	11	74	175	1461
20-25	808	32	229	699	44	316	35	109	289	2561
25-30	789	26	243	648	52	268	38	109	226	2399
30-35	916	17	257	825	74	291	52	123	281	2836
35-40	1118	27	313	1278	87	293	49	134	268	3567
40-45	926	13	250	1273	89	299	53	78	198	3179
45-50	855	9	203	1381	71	347	68	103	190	3227
50-55	684	14	136	1282	87	229	62	63	146	2703
55-60	502	6	77	972	49	151	46	66	77	1946
60-65	516	5	74	1249	83	162	52	92	122	2355
65-70	513	8	31	1360	75	164	56	115	95	2417
70-75	425	5	21	1268	90	121	44	119	82	2175
75-80	266	4	9	866	63	78	30	79	34	1429
80-85	159	2	2	479	39	18	18	46	19	782
85-90	70	1	0	259	16	10	9	18	10	393
90+	18	0	1	76	4	2	4	6	2	113
Female										
10-15	28	0	3	20	0	1	0	10	6	68
15-20	353	2	11	81	6	15	2	43	47	560
20-25	540	4	20	111	24	9	9	78	67	862
25-30	454	6	27	125	33	26	7	86	75	839
30-35	530	2	29	178	42	14	20	92	78	985
35-40	688	5	44	272	64	24	14	98	110	1319
40-45	566	4	24	343	76	18	22	103	86	1242
45-50	716	6	24	447	94	13	21	95	88	1504
50-55	942	7	26	691	184	21	37	129	131	2168
55-60	723	3	14	527	163	14	30	92	92	1658
60-65	820	8	8	702	245	11	35	140	114	2083
65-70	740	8	4	785	271	4	38	156	90	2096
70-75	624	6	4	610	244	1	27	129	46	1691
75-80	495	8	1	420	161	2	29	129	35	1280
80-85	292	3	2	223	78	0	10	84	23	715
85-90	113	4	0	83	14	0	6	34	2	256
90+	24	1	0	19	4	0	2	7	0	57
Sum	17565	253	2154	20377	2649	3118	937	2845	3313	53211

* asolliq: ingestion of solid or liquid matter; bgas: gas poisoning at home; cothergas: poisoning by other gas; dhss: hanging, strangling and suffocation; edrown: drowning; fgunexp: guns or explosives; gknives: knives, etc.; hjump: jumping; iother: other methods

The estimated parameters from the LBA are shown in Table S3. Latent budget 1 is mainly for younger adults from 15 to 50 years old and characterized by death from solid or liquid matter, poisoning of gases, gun or explosives and other methods. Latent budget 2 is composed of elderly females (more than 50 years old) who were dead from solid or liquid matter, drowning and jumping. Latent budget 3 contains mainly males (over almost all age ranges) who died from hanging, strangling or suffocation, gun or explosives and knives.

Table S3: Parameter Estimates from the Latent Budget Analysis of Example 3

Mixing parameters	**k = 1**	**k = 2**	**k = 3**	
female	0.10	0.90	0.00	
male	0.46	0.01	0.52	
10-15	0.15	0.00	0.85	
15-20	0.56	0.09	0.36	
20-25	0.66	0.07	0.26	
25-30	0.65	0.10	0.26	
30-35	0.59	0.15	0.27	
35-40	0.52	0.18	0.30	
40-45	0.45	0.17	0.38	
45-50	0.41	0.21	0.38	
50-55	0.28	0.38	0.34	
55-60	0.23	0.46	0.31	
60-65	0.15	0.51	0.35	
65-70	0.08	0.55	0.37	
70-75	0.02	0.64	0.35	
75-80	0.00	0.67	0.32	
80-85	0.00	0.72	0.28	
85-90	0.01	0.60	0.38	
90+	0.01	0.51	0.48	
budget propotion	0.39	0.35	0.27	
Latent budgets	**k = 1**	**k = 2**	**k = 3**	p_{+j}
asolliq	0.56	0.44	0.00	0.33
bgas	0.01	0.00	0.00	0.00
cothergas	0.12	0.00	0.00	0.04
dhss	0.00	0.32	0.83	0.38
edrown	0.00	0.10	0.05	0.05
fgunexp	0.11	0.00	0.07	0.06
gknives	0.01	0.02	0.03	0.02
hjump	0.05	0.08	0.02	0.05
iother	0.14	0.05	0.00	0.06

The corresponding qualitative evaluation of LBA-NN is shown in an importance plot in Figure S1. Notably, in the first panel, the death from solid or liquid matter is mainly for younger adults, reflected by higher importance values. Similar patterns are observed in death from all kinds of gas, gun or explosives and other methods. The females shows higher importance in the causes such as solid or liquid matter, drowning and jumping. On the contrary, males contribute more to causes including all kinds of gas, hanging, strangling or suffocation, gun or explosives and knives. The biplot derived from the importance table is shown in Figure S2. Note that three clusters are predefined so as to compare with the LBA (with three latent budgets) on the same level. Cluster 1 is characterized by death from all kinds of gas and gun or explosives. Cluster 3 is only composed of death from hanging, strangling or suffocations.

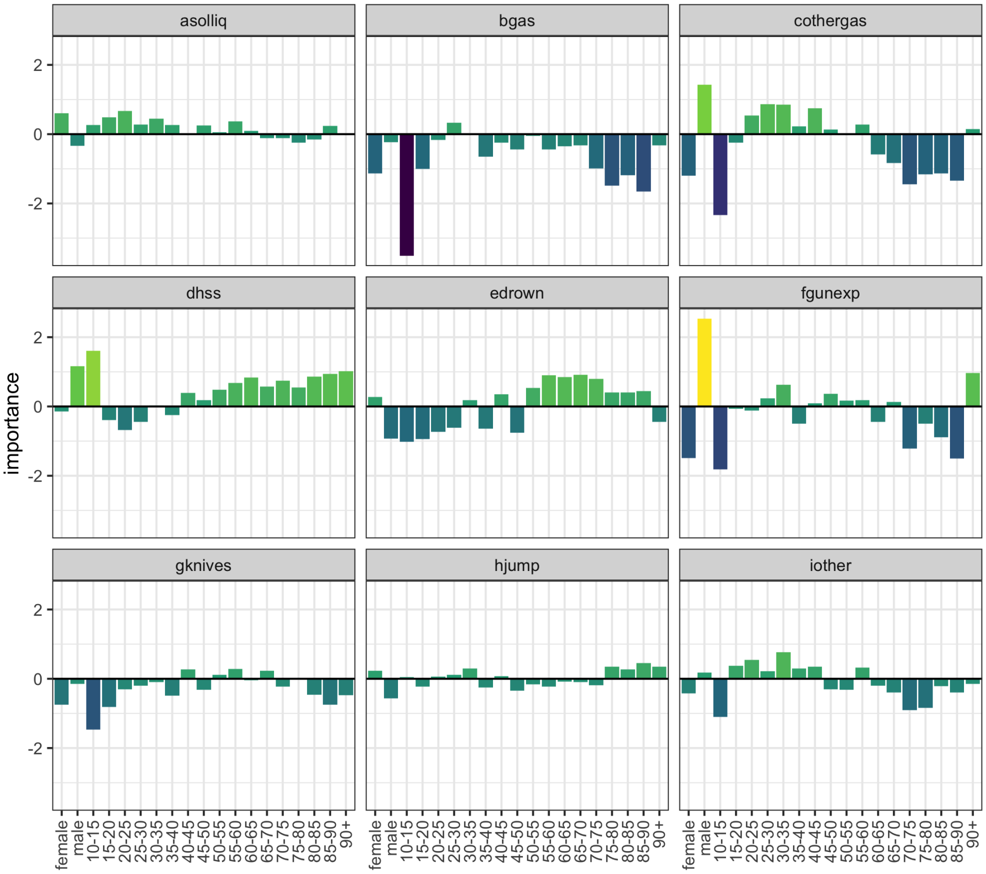


Figure S1: The importance plots of example 3 from LBA-NN. (x axis: categories of 2 variables, gender and age; y axis: importance value; each grid: for the categories of the response variable, cause of death)

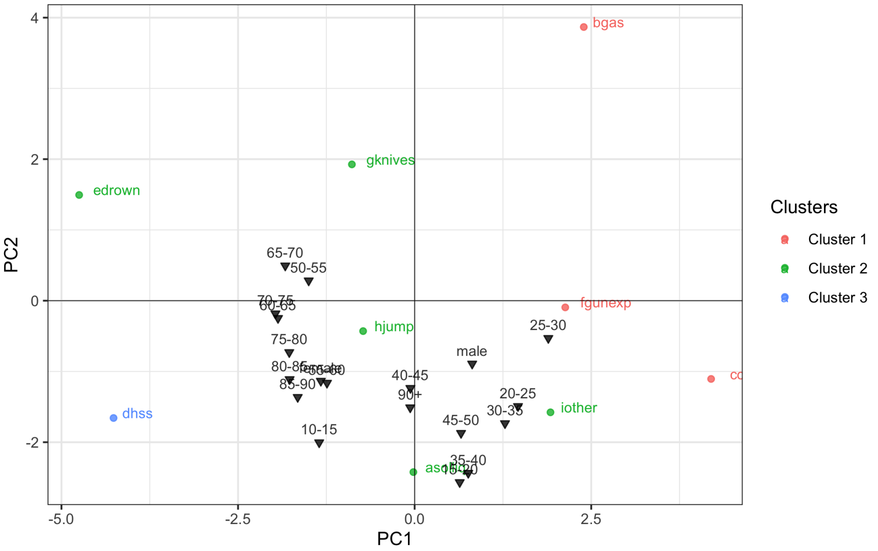
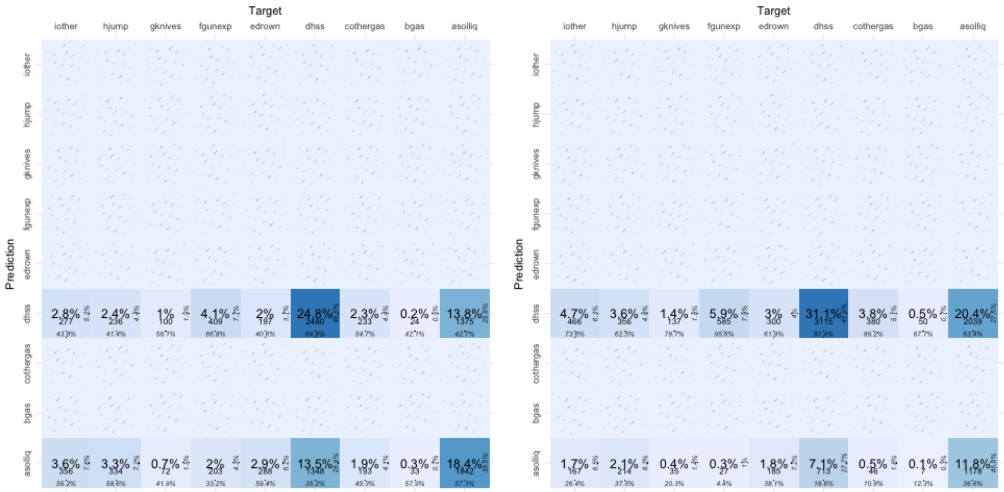


Figure S2: The biplot of example 3 from LBA-NN. (three clusters were predefined; explanatory categories presented in inverted triangle; response categories presented in points)

Apart from the qualitative evaluation, both models were implemented on the test dataset to compare their predictive abilities. The confusion matrices of the predictions from both models are shown in Figure S3. Both LBA and LBA-NN failed to predict causes other than solid or liquid matter and hanging, strangling or suffocation. The quantitative indicators are summarized in Table 8. Both models show similar performance in prediction. Both models have the accuracy of 0.43 and specificity of 0.90. LBA-NN has recall of 0.14 while LBA has that of 0.13. Only the mean square error shows relatively larger differences between LBA-NN and LBA (0.08 versus 0.11).



(a) (b)

Figure S3: The confusion matrices of LBA-NN and the LBA for example 3 (a: LBA-NN; b: LBA)

Table S4: Summary of predicative abilities of the two models for example 3

	LBA-NN	**LBA**
mean square error	0.08	0.11
accuracy	0.43	0.43
precision	-	-
recall	0.14	0.13
specificity	0.90	0.90
f1-score	-	-

Note:
Recall is also named as sensitivity. Since both LBA-NN and the LBA fail to predict classes other than “asolliq” and “dhss”, the precision and f1-score for LBA and LBA-NN are NA.

In this example, LBA-NN shows the same bad performance as LBA. Both methods were only able to predict two categories in the response variable. There are two underlying reasons. Firstly, group imbalance is obvious in the original dataset. The vast majority of the observations died from two categories, solid or liquid matter and hanging, strangling or suffocation. With the existence of the group imbalance, LBA-NN prioritized observations dead from the two causes during optimization to ensure the relatively low mean square error. In addition, unlike the other two examples, the explanatory variables in example 3 has in essence lower predictive ability. Intuitively, it can be extremely hard to predict causes of death only based on gender and age.

However, LBA-NN still produces fairly similar interpretation to LBA. For instance, cluster 1 composed of the causes including all kinds of gas, gun or explosives and other methods is similar to latent budget 1 from LBA.

Cluster 3 composed of the death from hanging, strangling or suffocation is similar to latent budget 3 from LBA.

Reference

Van der Heijden, P. G. M, A. Mooijaart, and J. De Leeuw. 1992. "Constrained Latent Budget Analysis." *Sociological Methodology* 22: 279. https://doi.org/10.2307/270999.