Finalised versions of normalisation and threshold

Contents

1 Idea

Even after excluding clashes, the list of harmonies left could be large and overwhelming for human consumption. Hence, there is a need to rank the harmonies basis how well they capture the variation in the measured variable and additionally reduce the number of harmonies for further exploration/visualization. Gestalt theory suggests that when items are placed in close proximity, people assume that they are in the same group because they are close to one another and apart from other groups. Hence, displays that capture more variation within different categories in the same group would be important to bring out different patterns of the data. Thus the idea here is to rate a harmony pair higher if this variation between different levels of the x-axis variable is higher on an average across all levels of facet variables.

2 Computing distances

One of the potential ways to evaluate this variation is by computing the pairwise distances between the distributions of the measured variable. We do this through Jensen-Shannon distance which is based on Kullback-Leibler divergence. Probability distributions are represented through sample quantiles instead of kernel density estimate so that there is minimal dependency on selecting kernel or bandwidth.

We shall call this measure of variation as Median Maximum Pairwise Distances (MMPD)

3 Normalize distances

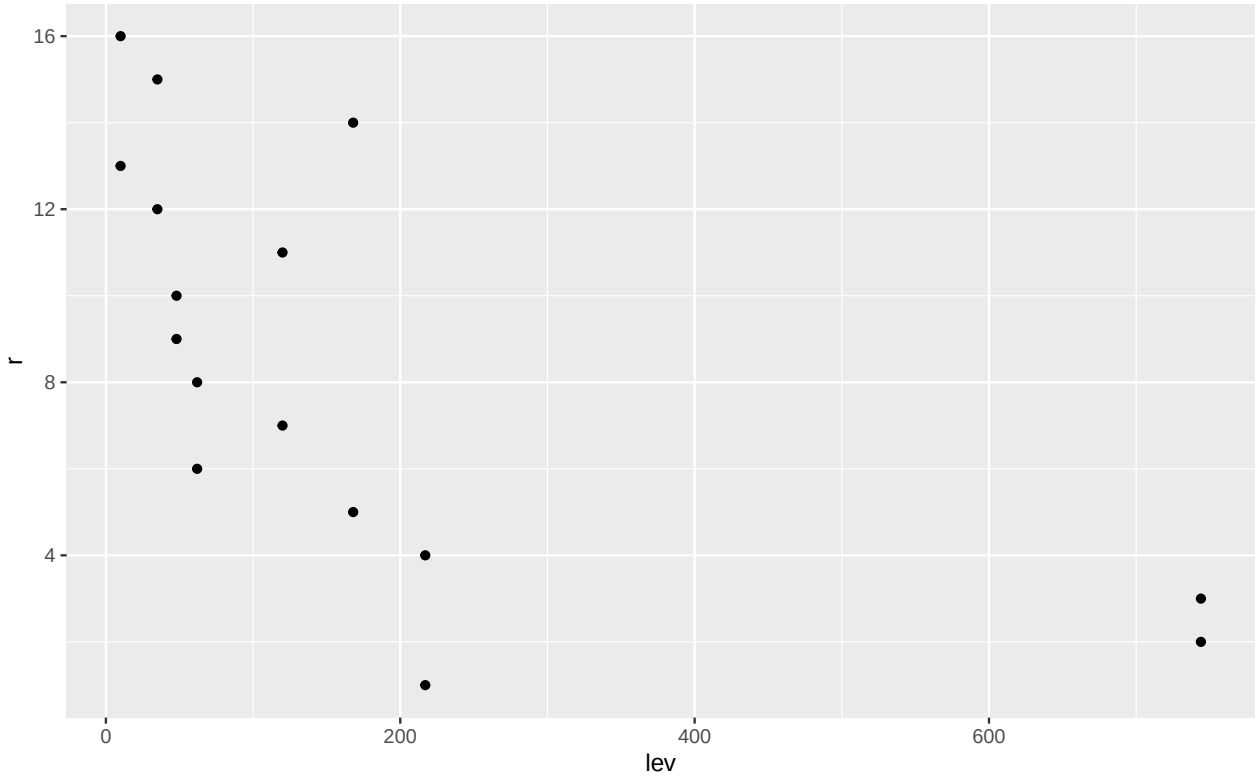
The harmony pairs could be arranged from highest to lowest average maximum pairwise distances across different levels of the harmonies. But maximum is not robust to the number of levels and is higher for harmonies with higher levels. Thus these maximum pairwise distances need to be normalized for different harmonies in a way that eliminates the effect of different levels. The Fisher–Tippett–Gnedenko theorem in the field of Extreme Value Theory states that the maximum of a sample of iid random variables after proper re- normalization can converge in distribution to only one of Weibull, Gumbel or Freschet distribution, independent of the underlying data or process. The normalizing constants, however, vary depending on the underlying distribution and hence it is important to assume a distribution of distances in our case.

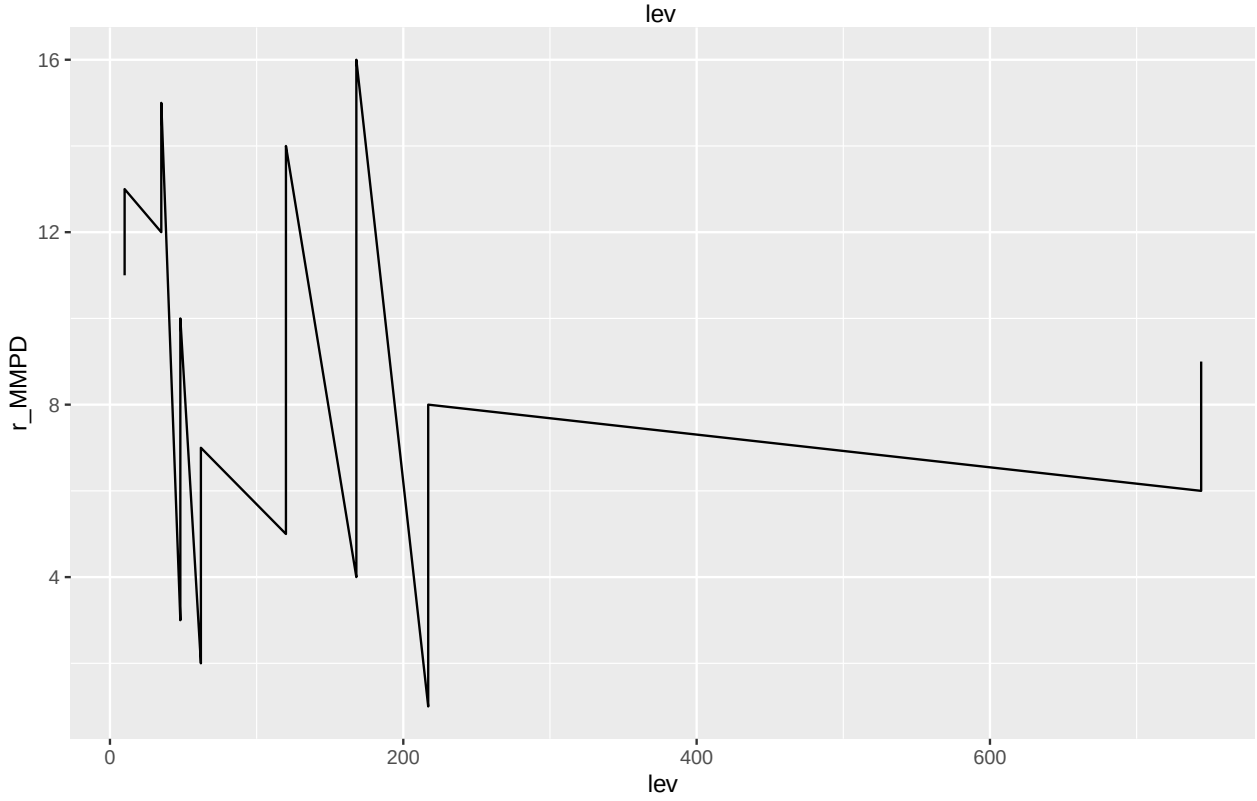
facet_variable	x_variable	facet_levels	x_levels	MMPD	max_pd	rankun	rankn
day_week	day_month	7	31	0.064	0.491	1	1
wknd_wday	day_month	2	31	0.060	0.048	8	2
wknd_wday	hour_day	2	24	0.044	0.041	9	3
day_week	hour_day	7	24	0.024	0.110	5	4
week_month	hour_day	5	24	0.023	0.050	7	5
hour_day	day_month	24	31	0.016	0.248	2	6
day_month	wknd_wday	31	2	0.014	0.069	6	7
day_month	day_week	31	7	0.011	0.115	4	8
day_month	hour_day	31	24	0.009	0.168	3	9
hour_day	wknd_wday	24	2	0.009	0.035	10	10
week_month	wknd_wday	5	2	0.007	0.021	13	11
day_week	week_month	7	5	0.004	0.022	12	12
wknd_wday	week_month	2	5	0.003	0.003	16	13
hour_day	week_month	24	5	0.003	0.026	11	14
week_month	day_week	5	7	0.001	0.003	15	15
hour_day	day_week	24	7	0.001	0.017	14	16

facet_variable	x_variable	facet_levels	x_levels	MMPD	max_pd	rankun	rankn
lag_field	over_match	2	40	0.490	0.603	1	1
inning_match	over	2	20	0.264	0.260	3	2
lag_field	over	2	20	0.182	0.188	6	3
inning_match	lag_field	2	2	0.090	0.069	8	4
lag_field	inning_match	2	2	0.060	0.065	9	5
over	lag_field	20	2	0.058	0.368	2	6
over	inning_match	20	2	0.039	0.201	5	7

4 Does normalisation work?

4.1 Number of levels and ranking





4.2 Comparing levels using simulated data

- MMPD should help choose the significantly different distributions only
- The ranking should be from most different to least different

4.3 How does it compare with taking unnormalised maximums?

5 Choose thresholds for harmonies

Permutation test:

Assumption: random permutation without considering ordering (global)

1. Given the data; $\{v_t : t = 0, 1, 2, \ldots, T-1\}$, the MMPD is computed and is represented by $MMPD_{obs}$.
2. From the original sequence a random permutation is obtained: $\{v_t^* : t = 0, 1, 2, \ldots, T-1\}$.
3. MMPD is computed for all random permutation of the data and is represented by $MMPD_{sample}$.
4. Steps (2) and (3) are repeated a large number of times M (e.g. 1000).
5. For each permutation, one $MMPD_{sample}$ value is obtained.
6. 95^{th} percentile of this $MMPD_{sample}$ distribution is computed and stored in $MMPD_{threshold}$.
7. If $MMPD_{obs} > MMPD_{threshold}$, harmony pairs are accepted. Onlyone threshold for all harmony pairs.

Pros: Considering thresholds global for all harmony pairs would imply less computation time.

Cons: Only one threshold for all harmony pairs might not be an appropriate measure but a good benchmark.