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. Assuming a numeric response variable, our graphics are displays of distributions compared across combinations of categorical variables, one placed at x-axis and the other on the facet. 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. Here, we have two main objectives:

- To choose harmonies for which distributions of categories are significantly different
- To rank the selected harmonies from highest to lowest variation in the distribution of their categories. 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.

The Jensen-Shanon distance between two probability distribution p_1 and p_2 is given by

$$d = [D(p_1, r) + D(p_2, r)]/2 \quad where \quad r = (p_1 + p_2)/2$$

where,

$$D(p_1, p_2) = \int_{-\infty}^{\infty} p_1(x) log \frac{p_1(x)}{p_2(x)}\, dx$$

is the Kullback-Leibler divergence between p_1 and p_2. Probability distributions are estimated through quantiles instead of kernel density so that there is minimal dependency on selecting kernel or bandwidth.

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

2.1 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.

More formally, $d_1, d_2 \ldots, d_n$ be a sequence of independent and identically-distributed pairwise distances and $M_n = \max\{d_1, \ldots, d_n\}$. Then Fisher–Tippett–Gnedenko theorem (Haan and Ferreira 2007) suggests that if a sequence of pairs of real numbers (a_n, b_n) exists such that each $a_n > 0$ and $\lim_{m\to\infty} P\left(\frac{M_n - b_n}{a_n} \leq x\right) = F(x)$, where F is a non-degenerate distribution function, then the limit distribution F belongs to either the Gumbel, Fréchet or Weibull family. The normalizing constants (a_n, b_n) vary depending on the underlying distribution of the pairwise distances. Hence to normalize appropriately, it is important to assume a distribution of these distances.

2.2 Distribution of distances

Theoretical: JS distances are distributed as chi-squared with m df where we discretize the continuous distribution with m discrete values. Taking sample percentiles to approximate the integral would mean taking $m = 99$. With large m, chi-squared is asymptotically normal by the CLT. Thus, by CLT, ${\chi^2}_m \tilde{} \mathcal{N}(m, 2m)$, which would depend on the number of discretization used to approximate the continuous distribution. Then $b_n = 1 - 1/n$ quantile of the normal distribution and $a_n = 1/[n * \phi(b_n)]$ where ϕ is the normal density function. n is the number of pairwise comparisons being made.

Empirical: Distribution of JS distances is assumed to be normal but the mean and variance are estimated from the sample, rather than deducing it from the number of discretization used to approximate the continuous distribution.

3 Choose thresholds for harmonies through 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 means we are assuming distribution of all harmonies pairs are similar, which might not be the case.But nevertheless, it is a good benchmark.

```
#> # A tibble: 16 x 7
#>    facet_variable x_variable facet_levels x_levels    MMPD max_pd     r
#>    <chr>          <chr>             <int>    <int>   <dbl>  <dbl> <dbl>
#>  1 wknd_wday      day_month             2       31 3.27    0.0477    13
#>  2 wknd_wday      hour_day              2       24 1.94    0.0574    12
#>  3 wknd_wday      week_month            2        5 1.31    0.0126    16
#>  4 day_week       hour_day              7       24 0.849   0.0868     7
#>  5 day_week       day_month             7       31 0.652   0.251      3
#>  6 week_month     hour_day              5       24 0.560   0.113      5
#>  7 day_week       week_month            7        5 0.489   0.0625    11
#>  8 week_month     day_week              5        7 0.431   0.0637    10
#>  9 hour_day       day_month            24       31 0.301   0.214      4
#> 10 hour_day       week_month           24        5 0.279   0.0890     6
#> 11 day_month      day_week             31        7 0.250   0.253      2
#> 12 day_month      hour_day             31       24 0.223   0.269      1
#> 13 hour_day       day_week             24        7 0.147   0.0766     8
#> 14 day_month      wknd_wday            31        2 0.0136  0.0683     9
#> 15 hour_day       wknd_wday            24        2 0.00879 0.0301    14
#> 16 week_month     wknd_wday             5        2 0.00684 0.0208    15
```

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

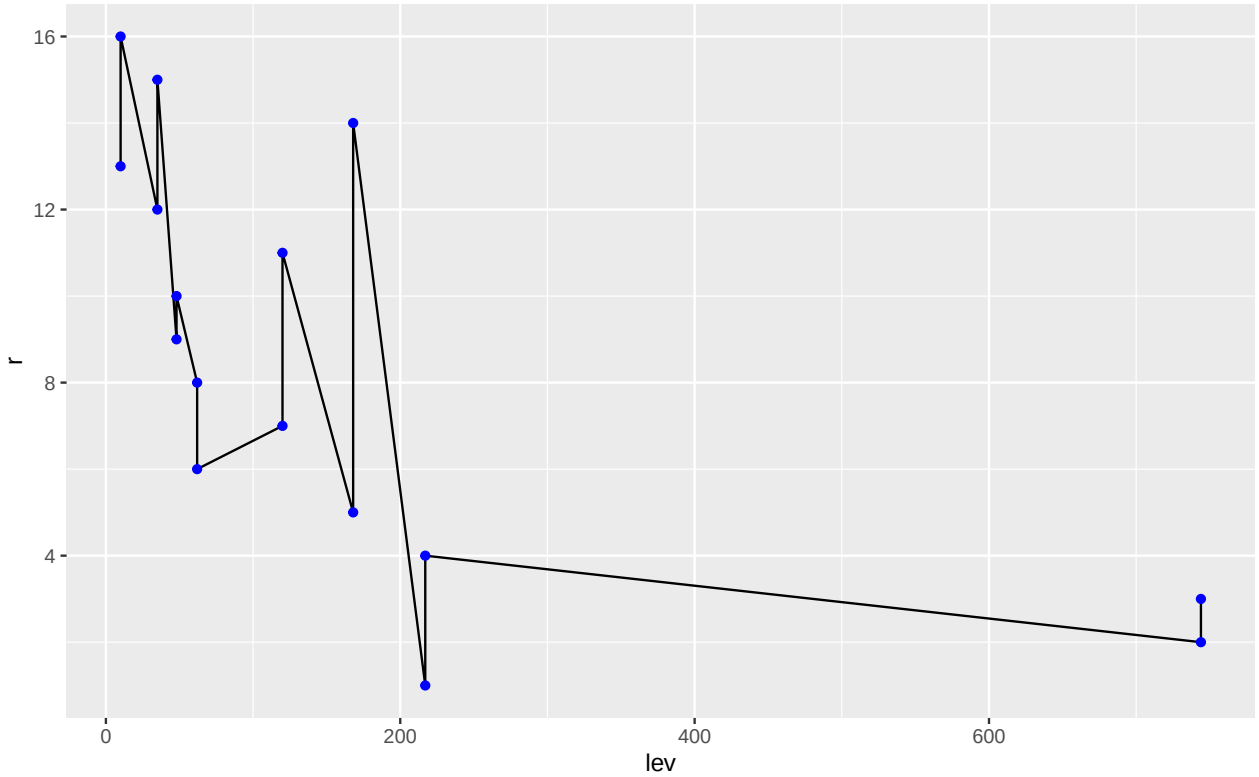
4 Does normalisation work?

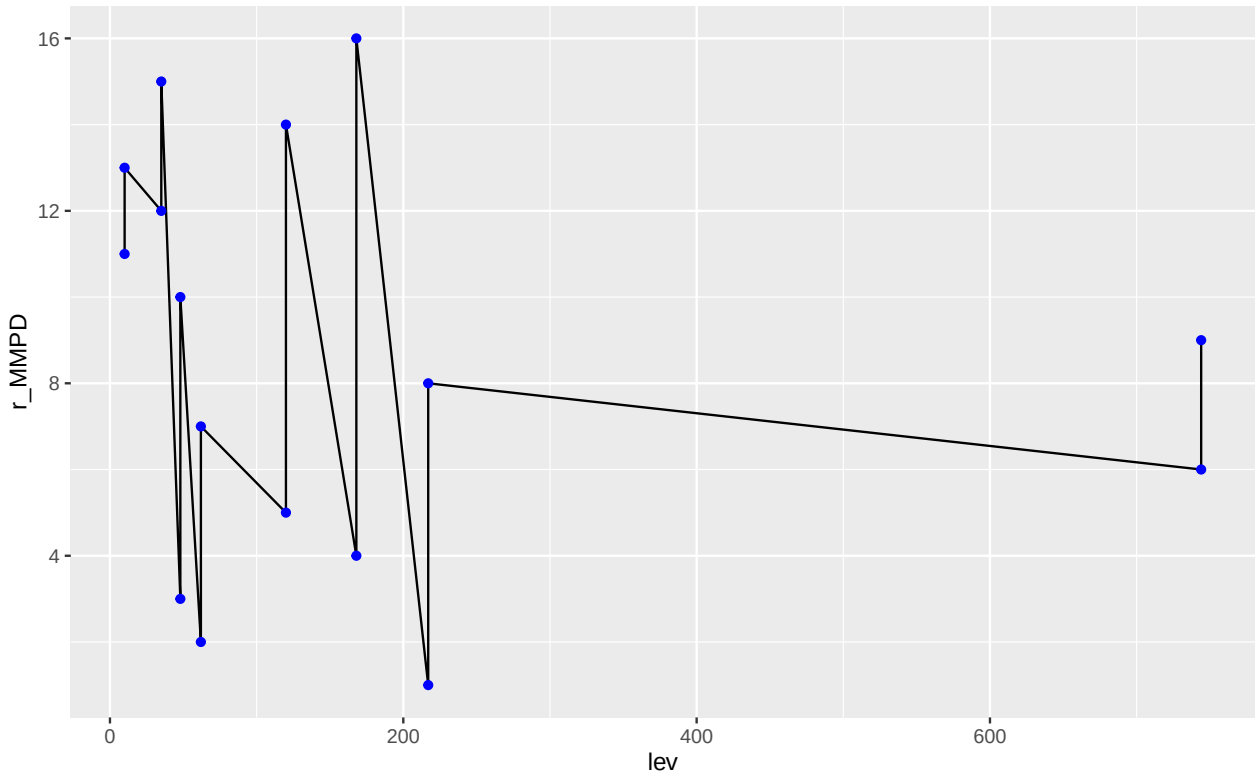
4.1 Minimal example

Consider three cyclic granularities A, B and C with 2, 3 and 4 categories. Thus, the harmony table consisting of all possible harmony pairs (assuming all pairs are harmonies), would like the following:

facet_variable	x_variable	facet_levels	x_levels
A	B	2	3
B	A	3	2
A	C	2	4
C	A	4	2
B	C	3	4
C	B	4	3

4.2 Number of levels and ranking





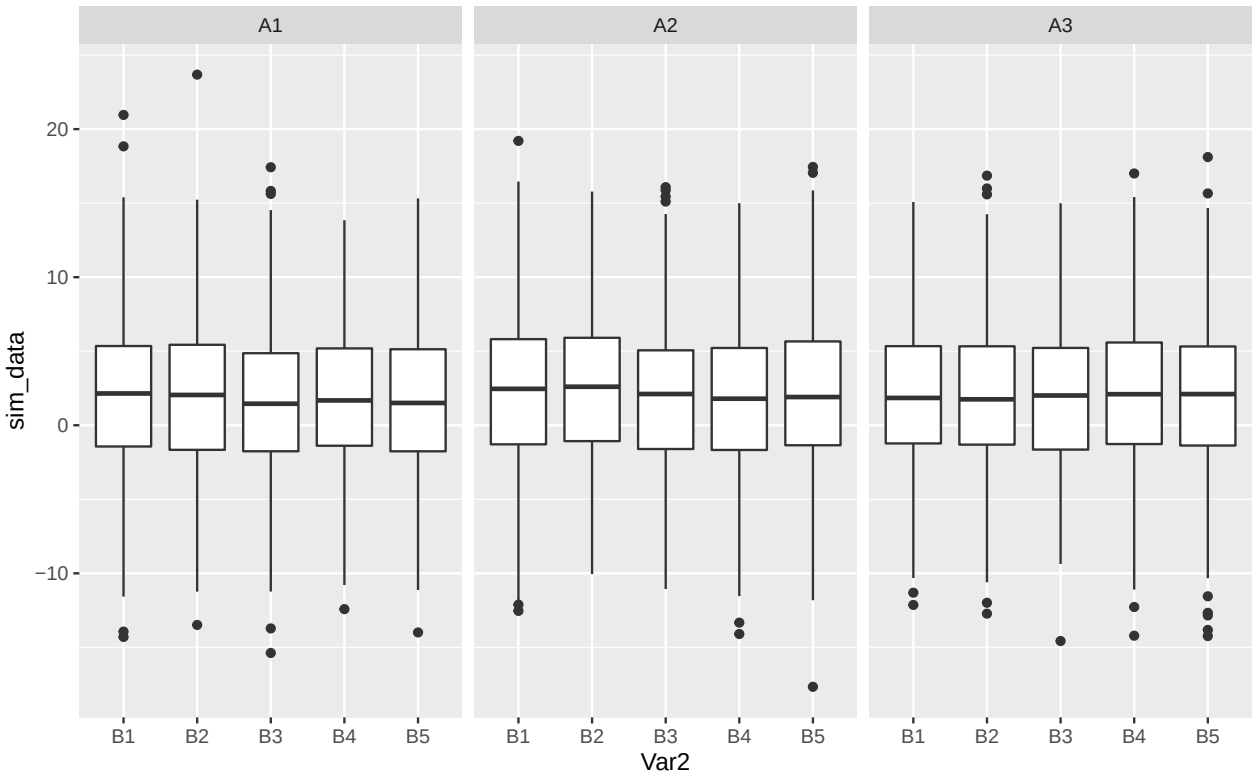
4.3 Comparing levels using simulated data

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

Both of these should be checked again taking unnormalised statistic

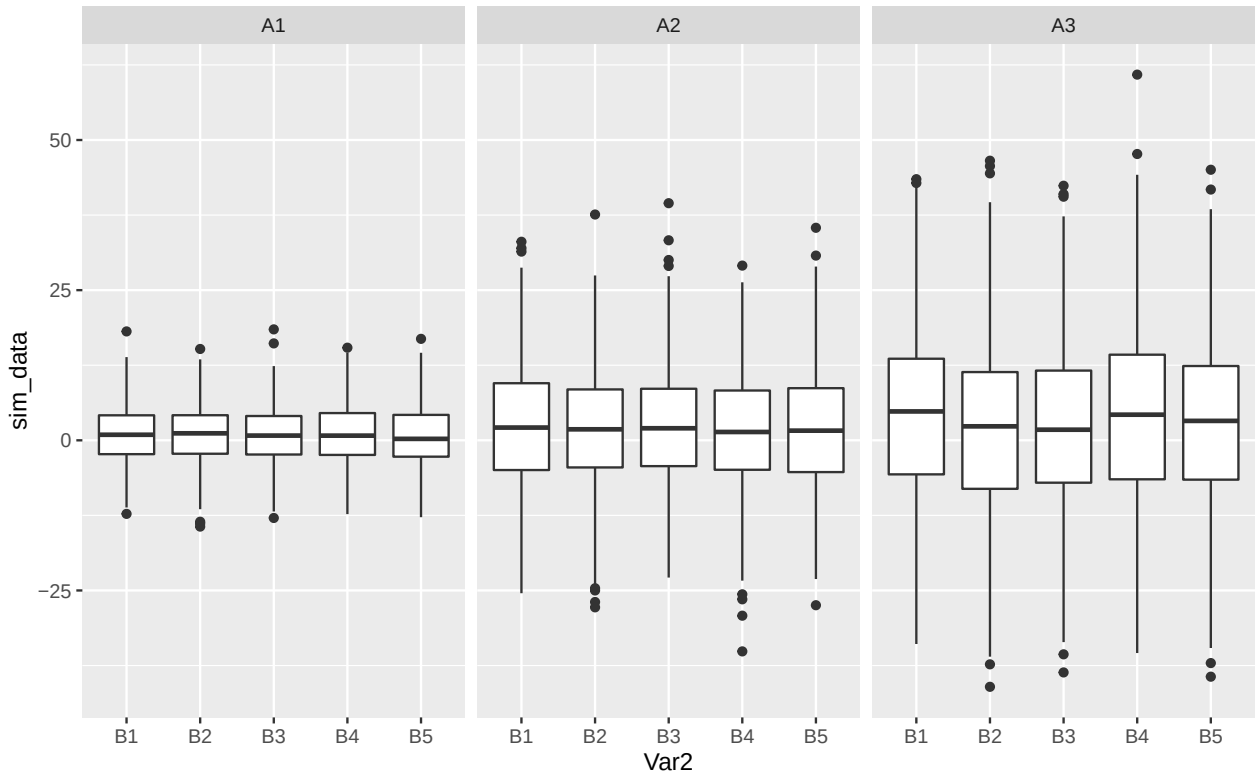
4.4 Scenario 1: Simulated same normal distributions for all combinations of the pair of cyclic granularities

facet_variable	x_variable	facet_levels	x_levels	MMPD	max_pd	r	gt_MMPD	gt_maxpd
A	B	3	5	0.43971	0.05075	1	FALSE	FALSE
C	B	7	5	0.43203	0.04804	3	FALSE	FALSE
D	E	24	31	0.22246	0.04729	4	FALSE	FALSE
D	C	24	7	0.16933	0.04948	2	FALSE	FALSE



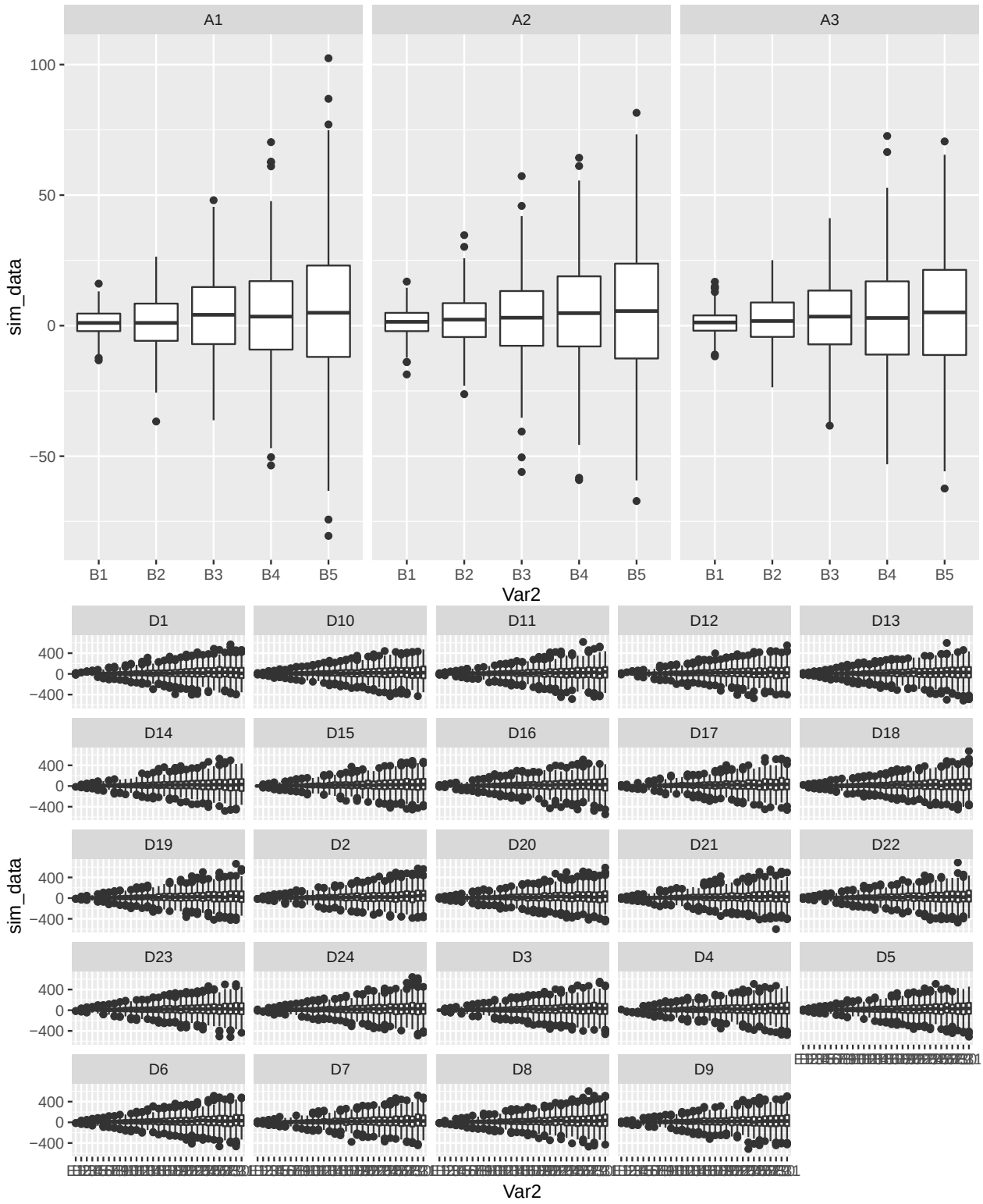
4.5 Scenario 2: Simulated same normal distributions for all x-levels within a facet, but different distributions across facets

facet_variable	x_variable	facet_levels	x_levels	MMPD	max_pd	r	gt_MMPD	gt_maxpd
A	B	3	5	0.49000	0.04612	2	TRUE	FALSE
C	B	7	5	0.33748	0.04531	3	FALSE	FALSE
D	E	24	31	0.25856	0.05473	1	FALSE	TRUE
D	C	24	7	0.15213	0.04260	4	FALSE	FALSE



4.6 Scenario 3: Simulated different normal distributions for different x-levels but they do not vary across facets

facet_variable	x_variable	facet_levels	x_levels	MMPD	max_pd	r	gt_MMPD	gt_maxpd
D	E	24	31	1.84640	0.12771	4	TRUE	FALSE
A	B	3	5	0.79410	0.13663	2	TRUE	FALSE
C	B	7	5	0.48567	0.12987	3	FALSE	FALSE
D	C	24	7	0.45191	0.15607	1	FALSE	FALSE


A1
A2
A3
100
50
0
−50
sim_data
B1
B2
B3
B4
B5
Var2
D1
D10
D11
D12
D13
D14
D15
D16
D17
D18
D19
D2
D20
D21
D22
D23
D24
D3
D4
D5
D6
D7
D8
D9
400
0
−400
sim_data
Var2

Scenario 4: Mixture of secenario 2 and 3

Haan, Laurens de, and Ana Ferreira. 2007. *Extreme Value Theory: An Introduction*. Springer Science & Business Media.