Screening harmonies

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 divergence 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.

3.1 Theoretical evidence

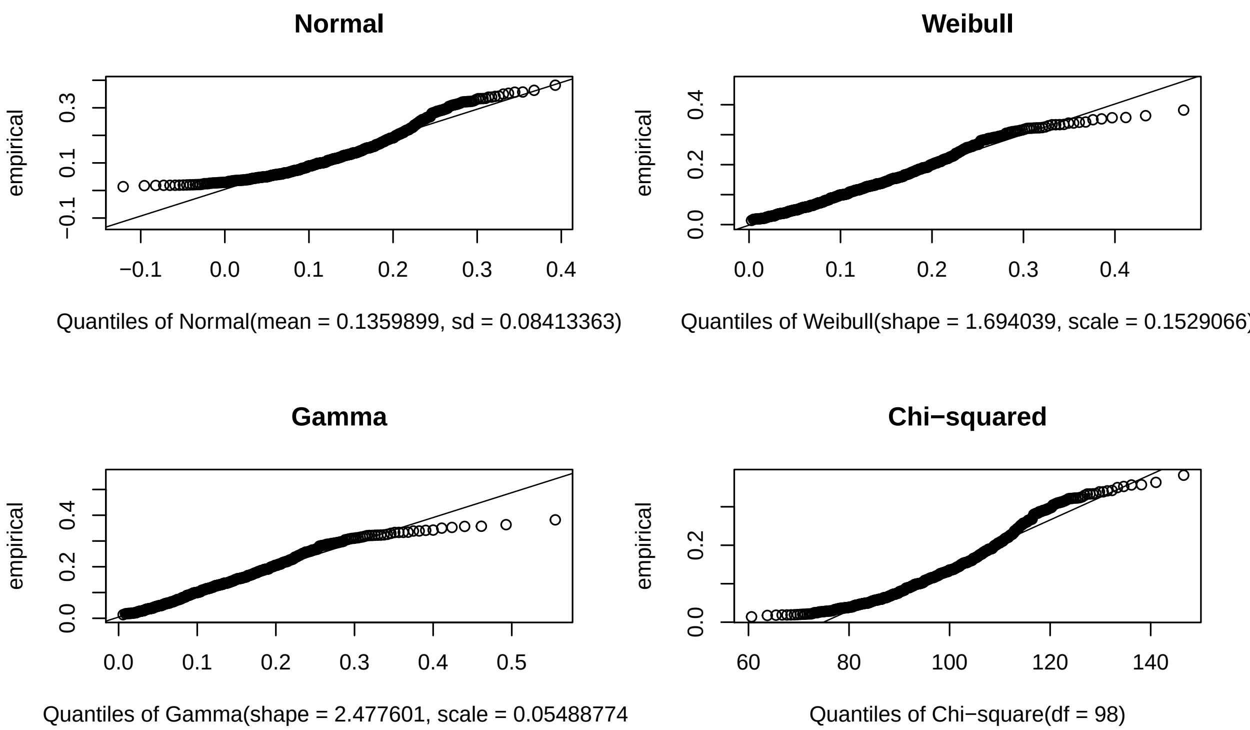
Menéndez et al. (1997) and Grosse et al. (2002) provide studies of the statistical properties of the Jensen-Shannon divergences and suggest that the theoretical asymptotic distribution of Jensen-Shannon divergence converges to a Chi-squared distribution. Let $p^{(1)} \equiv (p_1^{(1)}, p_2^{(1)}, \ldots, p_k^{(1)})$ and $p^{(2)} \equiv (p_1^{(2)}, p_1^{(2)}, \ldots, p_k^{(2)})$ denote two probability distributions, where k is the number of components of the probability vector p. Then 2N(ln2)D is known to converge to $\{\chi^2\}$ distribution with degrees of freedom = (k - 1) and

$D[p^{(1)}, p^{(2)}] = H((p^{(1)} + p^{(2)})/2) - H(p^{(1)}) - H(p^{(2)})$, $H_{(p)} = \sum_{i=1}^{k} p_i log_2 p_i$ and N is the total number of events.

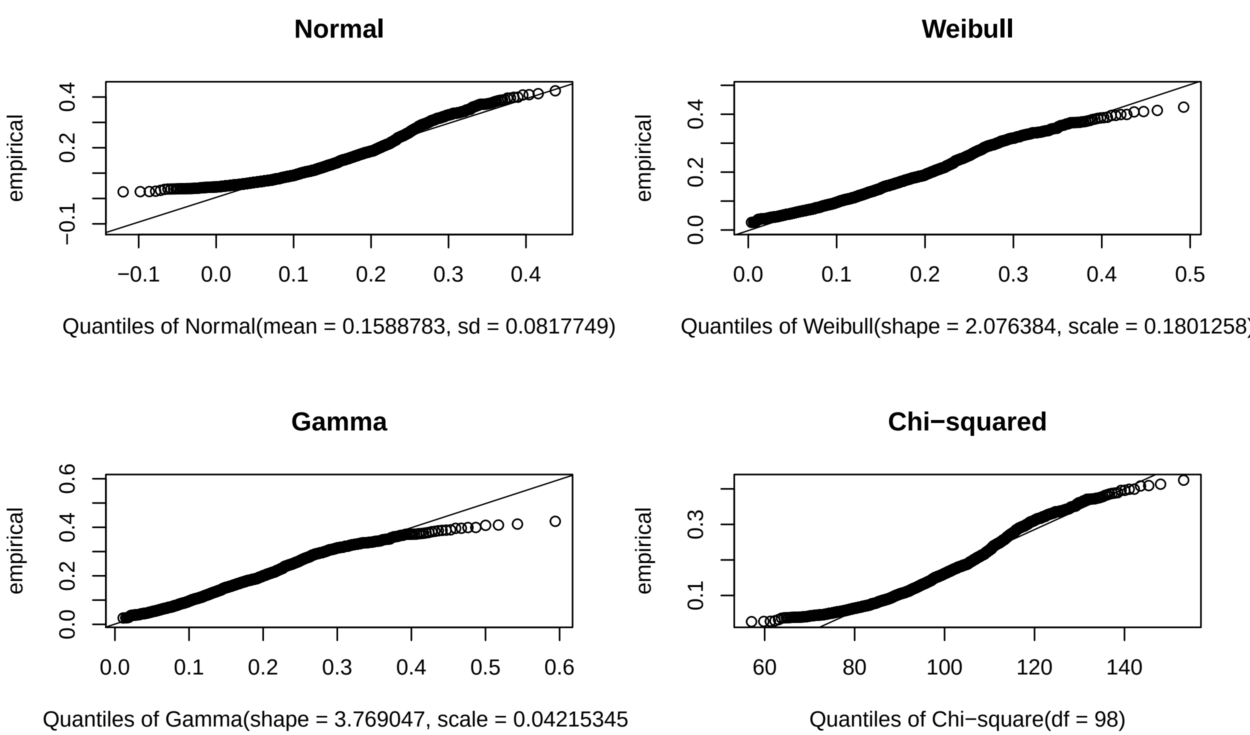
3.2 Empirical evidence

However, since we are dealing with small finite samples, a more appropriate approach would be to look at the distribution of the samples through histogram, density plots or using QQ-plot to see how well the empirical quantiles match the theoretical quantiles. The QQ plots of four harmony pairs are plotted below. It could be seen that Chi-square distribution serves as a pretty good fit to the data (specially in the extreme right tail).

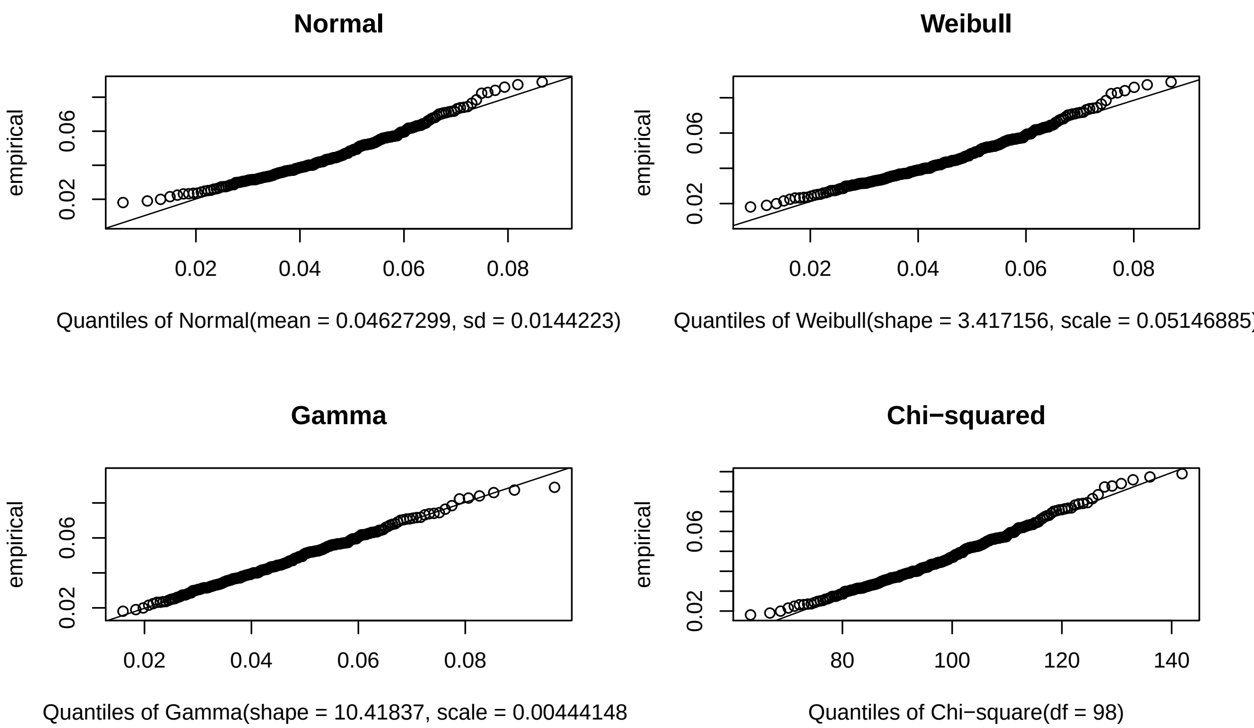
3.2.1 Distribution fitting of distances for the harmony pair (weekend/weekday, hour-of-day)



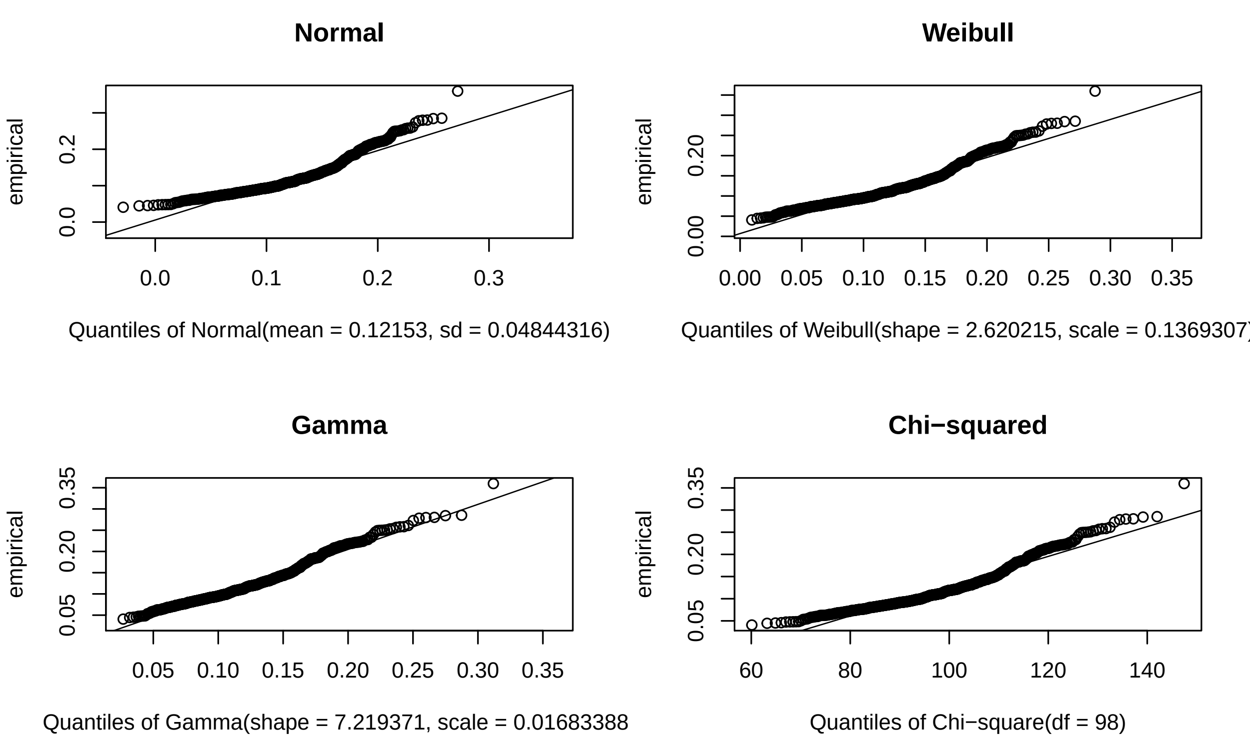
3.2.2 Distribution fitting of distances for the harmony pair (day-of-week, hour-of-day)



3.2.3 Distribution fitting of distances for the harmony pair (hour-of-day, week-of-month)



3.2.4 Distribution fitting of distances for the harmony pair (day-of-month, day-of-week)



4 Estimating parameters

Currently, MME is used. But MASS::fitdistr() and package fitdistrplus also provide methods to estimate parameters through MLE. WIP.

5 Choose thresholds for harmonies

Threholds could be chosen for distances (chi-squared) or maximum distance (Gumbel distribution).

Critical values of the Chi-squared statistic could be obtained for appropriate degrees of freedom.

Chi-squared statistic: And test values less than critical value would imply that the distances are not significantly different from zero , implying distributions are similar. If distributions are similar most times, then the plot is not interesting. Hence, all pairs for which most of the pairwise distances are significantly different from zero would only be included in the harmony rank table.

Gumbel statistic: The test decides if the sample of maximum distance is from Gumbel or not. If it is from Gumbel, the value of the test statistic should be ideally zero. And we want to take all pairs for which the statistic is significantly different from zero. So we choose test values greater than the Gumbel critical value at 5% significance.

6 Results

6.1 Smart meter data

6.1.1 Maximum distance between levels of x-axis variable and median across levels of facet variable

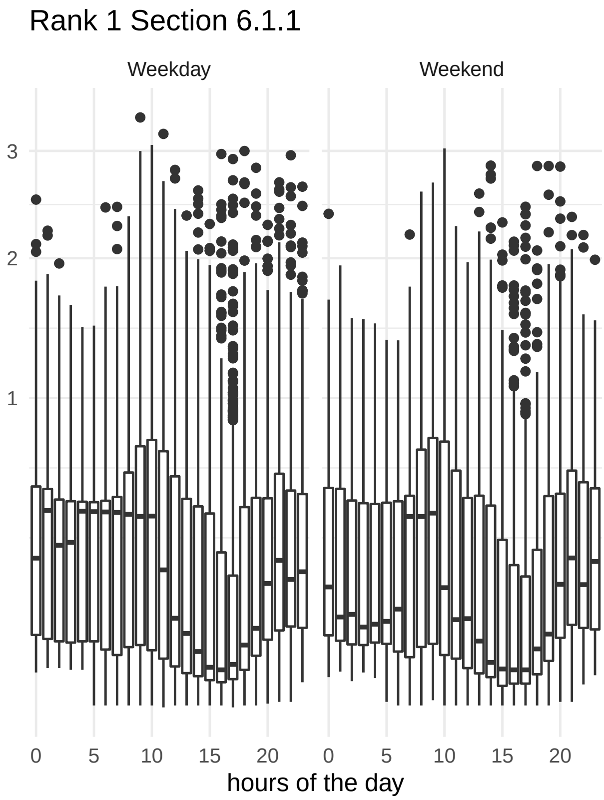
facet_variable	x_variable	f_L	x_L	chi	weibull	gamma	normal	general
wknd_wday	hour_day	2	24	1	1	1	1	1
week_month	hour_day	5	24	2	4	3	2	2
day_week	hour_day	7	24	3	7	6	4	3
day_week	day_month	7	31	4	5	8	3	5
day_month	hour_day	31	24	5	12	7	6	7
wknd_wday	day_month	2	31	6	3	11	5	4
hour_day	day_month	24	31	7	13	10	7	10
day_month	day_week	31	7	8	9	9	8	11
day_month	wknd_wday	31	2	9	14	2	14	16
hour_day	wknd_wday	24	2	10	15	4	15	15
week_month	wknd_wday	5	2	11	16	5	16	13
day_week	week_month	7	5	12	6	12	12	8
wknd_wday	week_month	2	5	13	2	16	11	6
hour_day	day_week	24	7	14	11	14	9	14
hour_day	week_month	24	5	15	10	13	13	12
week_month	day_week	5	7	16	8	15	10	9

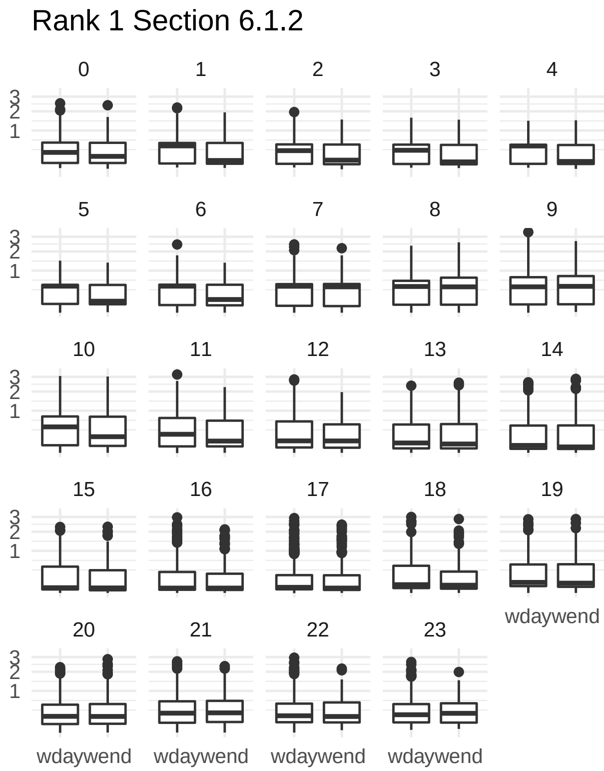
facet_variable	x_variable	rank_chi	rank_weibull	rank_gamma	rank_normal	rank_general	thresval
wknd_wday	hour_day	1	1	1	1	1	select
week_month	hour_day	2	4	3	2	2	select
day_week	hour_day	3	7	6	4	3	select
day_week	day_month	4	5	8	3	5	select
day_month	hour_day	5	12	7	6	7	select
wknd_wday	day_month	6	3	11	5	4	NA
hour_day	day_month	7	13	10	7	10	select
day_month	day_week	8	9	9	8	11	select
day_month	wknd_wday	9	14	2	14	16	NA
hour_day	wknd_wday	10	15	4	15	15	NA
week_month	wknd_wday	11	16	5	16	13	NA
day_week	week_month	12	6	12	12	8	NA
wknd_wday	week_month	13	2	16	11	6	NA
hour_day	day_week	14	11	14	9	14	NA
hour_day	week_month	15	10	13	13	12	NA
week_month	day_week	16	8	15	10	9	NA

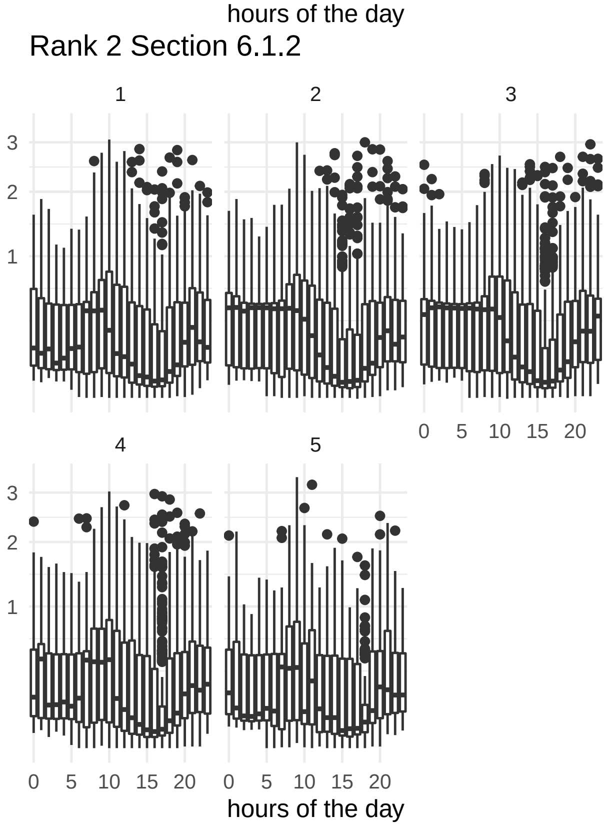
6.1.2 Maximum distance between levels of facet variable and median across levels of x-axis variable

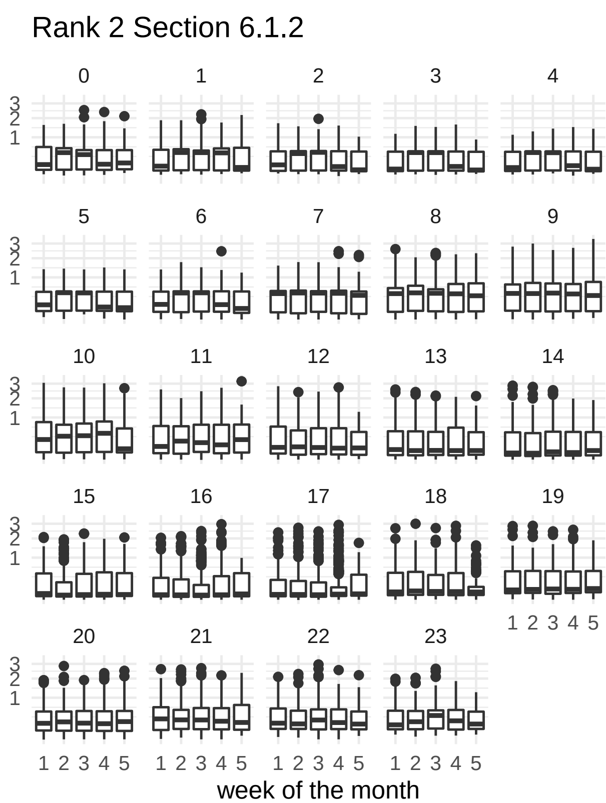
facet_variable	x_variable	facet_levels	x_levels	chi	weibull	gamma	normal	general
hour_day	wknd_wday	24	2	1	1	1	1	1
hour_day	week_month	24	5	2	4	3	3	2
hour_day	day_week	24	7	3	7	6	4	3
day_month	day_week	31	7	4	5	8	2	5
hour_day	day_month	24	31	5	12	7	6	7
day_month	wknd_wday	31	2	6	3	11	5	4
day_month	hour_day	31	24	7	13	10	7	10
day_week	day_month	7	31	8	9	9	8	11
wknd_wday	day_month	2	31	9	14	2	15	16
wknd_wday	hour_day	2	24	10	15	4	14	15
wknd_wday	week_month	2	5	11	16	5	16	13
week_month	day_week	5	7	12	6	12	13	8
week_month	wknd_wday	5	2	13	2	16	11	6
day_week	hour_day	7	24	14	11	14	9	14
week_month	hour_day	5	24	15	10	13	12	12
day_week	week_month	7	5	16	8	15	10	9

6.2 Graphical evidence



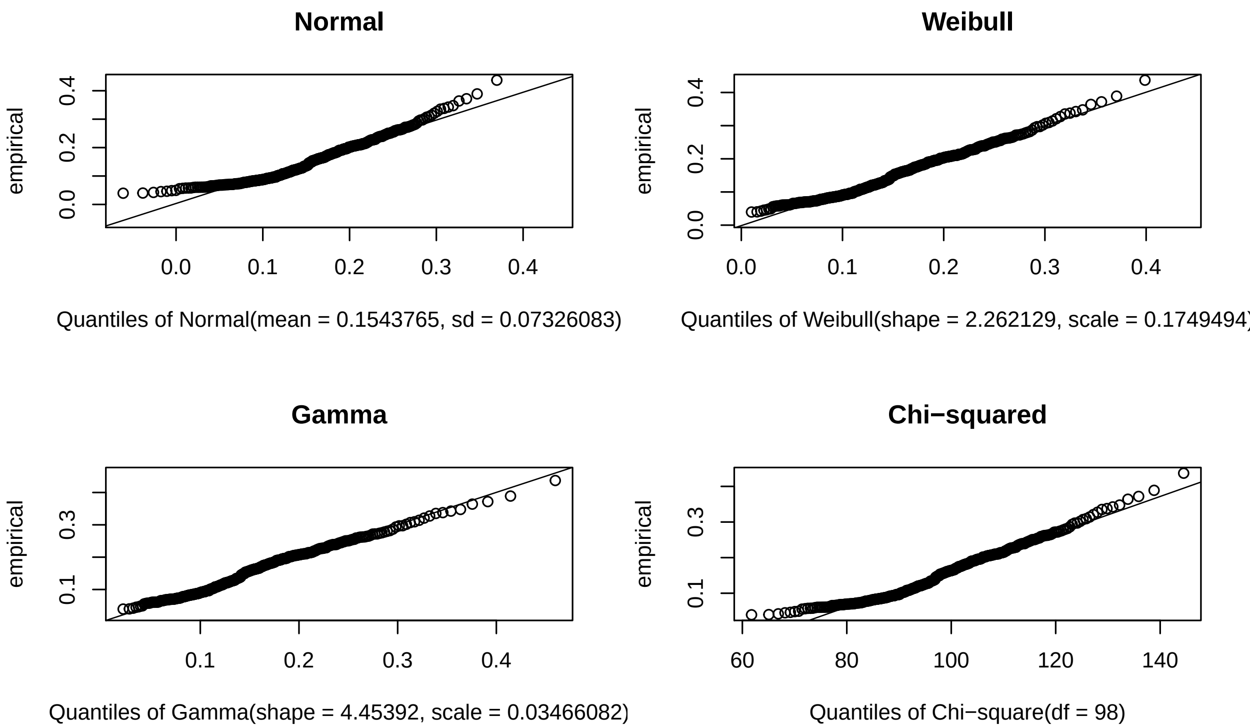






6.3 cricket data

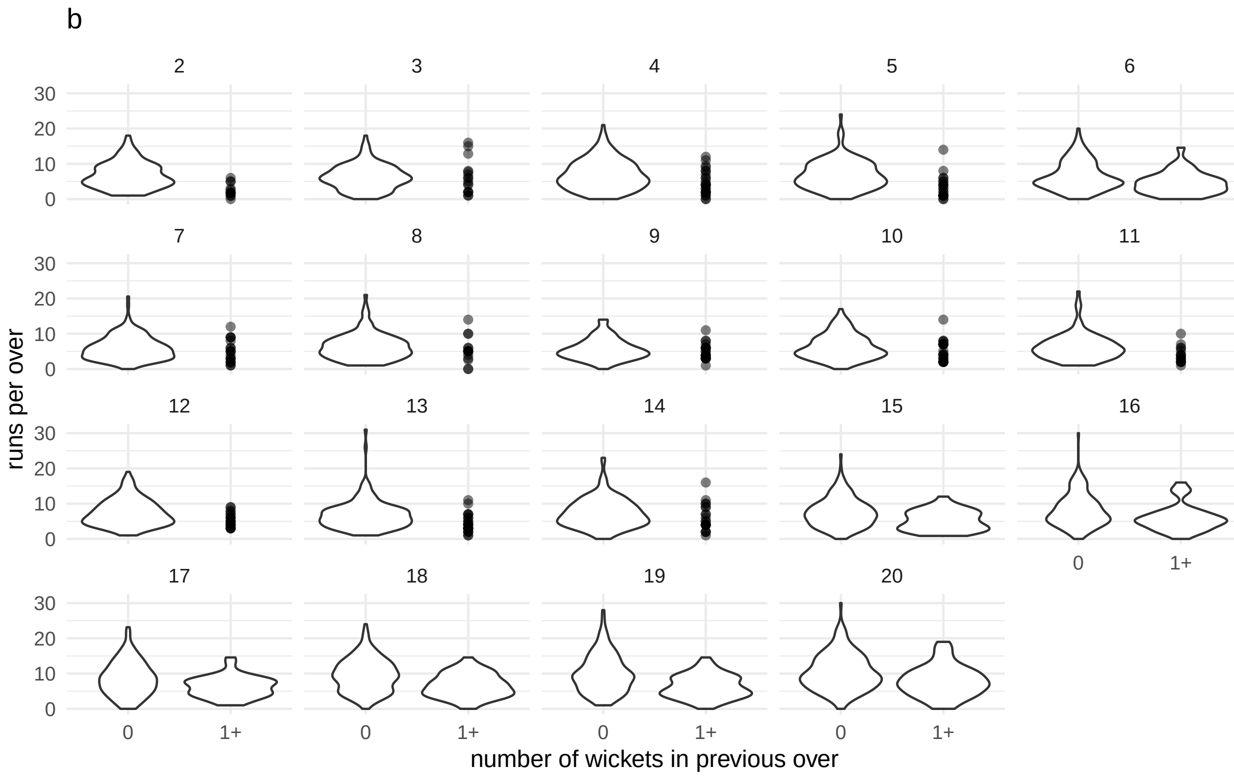
6.3.1 Distribution fitting of distances for the harmony pair (lag_field, over)



facet_variable	x_variable	facet_levels	x_levels	chi	weibull	gamma	normal	general
lag_field	over_match	2	40	1	4	7	1	1
lag_field	over	2	20	2	1	8	2	2
inning_match	over	2	20	3	2	9	3	3
lag_field	over_inning	2	20	4	3	10	4	5
inning_match	lag_field	2	2	5	6	1	6	7
inning_match	over_inning	2	20	6	5	11	5	4
lag_field	inning_match	2	2	7	7	2	7	6
over	lag_field	20	2	8	8	3	8	11
over_inning	lag_field	20	2	9	9	4	9	10
over	inning_match	20	2	10	10	5	10	9
over_inning	inning_match	20	2	11	11	6	11	8

6.3.2 Maximum distance between levels of facet variable and median across levels of x-axis variable

facet_variable	x_variable	facet_levels	x_levels	chi	weibull	gamma	normal	general
over	lag_field	20	2	1	1	8	1	1
over	inning_match	20	2	2	2	9	2	2
over_inning	lag_field	20	2	3	3	10	3	4
lag_field	inning_match	2	2	4	5	1	5	5
over_inning	inning_match	20	2	5	4	11	4	3
inning_match	lag_field	2	2	6	6	2	6	6
lag_field	over	2	20	7	7	3	7	10
lag_field	over_match	2	40	8	8	4	8	11
lag_field	over_inning	2	20	9	9	5	9	8
inning_match	over	2	20	10	10	6	10	9
inning_match	over_inning	2	20	11	11	7	11	7



Bibliography

Grosse, Ivo, Pedro Bernaola-Galván, Pedro Carpena, Ramón Román-Roldán, Jose Oliver, and H Eugene Stanley. 2002. "Analysis of Symbolic Sequences Using the Jensen-Shannon Divergence." *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 65 (4 Pt 1): 041905.

Menéndez, M L, J A Pardo, L Pardo, and M C Pardo. 1997. "The Jensen-Shannon Divergence." *J. Franklin Inst.* 334 (2): 307–18.