Supplementary materials for the main submission entitled - A new metric for automatic discovery of periodic patterns in time series

Contents

Appendix A recalls the necessary notations and

1 Recalling notations

Consider two cyclic granularities A and B, such that $A = \{a_j : j = 1, 2, \dots, J\}$ and $B = \{b_k : k = 1, 2, \dots, K\}$ with A placed across x-axis and B across facets. Let $v = \{v_t : t = 0, 1, 2, \dots, T-1\}$ be a continuous measured variable observed across T time points. Let the four elementary designs be D_{null} where there is no difference in distribution of v for A or B, D_{var_f} denotes the set of designs where there is difference in distribution of v for B and not for A. Similarly, D_{var_x} denotes the set of designs where difference is observed only across A. Finally, $D_{var_{all}}$ denotes those designs for which difference is observed across both A and B.

2 Simulation design

Observations are generated from Gamma(2,1), G(0.5, 1), N(0,1), N(0, 5) and N(5, 1) distribution for each combination of nx and $nfacet$ from the following sets: $nx = nfacet = \{2, 3, 5, 7, 14, 20, 31, 50\}$ to cover a wide range of levels from very low to moderately high. Each combination is being referred to as a *panel*. That is, data is being generated for each of the panels $\{nx = 2, nfacet = 2\}, \{nx = 2, nfacet = 3\}, \{nx = 2, nfacet = 5\}, \dots, \{nx = 50, nfacet = 31\}, \{nx = 50, nfacet = 50\}$. For each of the 64 panels, $ntimes = 500$ observations are drawn for each combination of the categories. That is, if we consider the panel $\{nx = 2, nfacet = 2\}$, 500 observations are generated for each of the combination of categories from the panel, namely, $\{(1, 1), (1, 2), (2, 1), (2, 2)\}$. The values of *wpd* is obtained for each of the panels. The measurement variable for each combination of categories in a panel are drawn from the same distribution and hence the design corresponds to D_{null} . Furthermore, this entire method is repeated for each panels $nsim = 200$ times, so that the distribution of *wpd* under D_{null} could be observed.

3 Behavior of raw weighted distance measure

Following the work of Krzysztofowicz (1997) the empirical NQT involves the following steps: 1. Sort the sample of measured variable X from the smallest to the largest observation $x_{(1)}, \ldots, x_{(i)}, .., x_{(n)}$. 2. Estimate the cumulative probabilities $p_{(1)}, \ldots, p_{(i)}, .., p_{(n)}$. using a plotting position like $i/(n+1)$ such that $p_{(i)} = P(X \leq x_{(i)})$. 3. Transforming each observation $x_(i)$ of X into observation $y(i) = Q^{-1}(p(i))$ of the standard normal variate Y , with Q denoting the standard normal distribution and Q^{-1} its inverse, applying discrete mapping.

Firstly, we show that NQT works to make the distribution within each panel same. Next, we study the distribution of *wpd* for different nx and $nfacet$ in various ways.

3.1 Normal quantile transform

Since the measure *wpd* is essentially set up to detect "differences" in distributions irrespective of underlying distribution, it would be ideal if it has minimal dependency on the type, location and scale of the initial distribution. To that end, some data pre-processing through the Normal Score Transform (NQT) has been applied in order to make most asymmetrical distributed measured variables more normal-like. Figure 1 shows ridge plots of raw *wpd* for a Gamma(0.5,1), Gamma(2,1) before NQT. It is observed that for the underlying distribution Gamma(2,1), location and scale of the distribution of wpd changes from top-left panel to bottom-right panel. Moreover, the location and scale of the distribution of *wpd* fir different underlying distribution Gamma(0.5,1), Gamma(2,1). Figure 2 shows the the distributions of *wpd* under same underlying distributions but after performing NQT. It is observed that within each panel, the distributions of the *wpd* looks same, however, the distributions change from extreme top-left panel to bottom-right panels. This implies, NQT has atleast been able to bridge the gap in distribution of *wpd* for different non-normal underlying distributions.

3.2 By distribution

3.3 By category levels

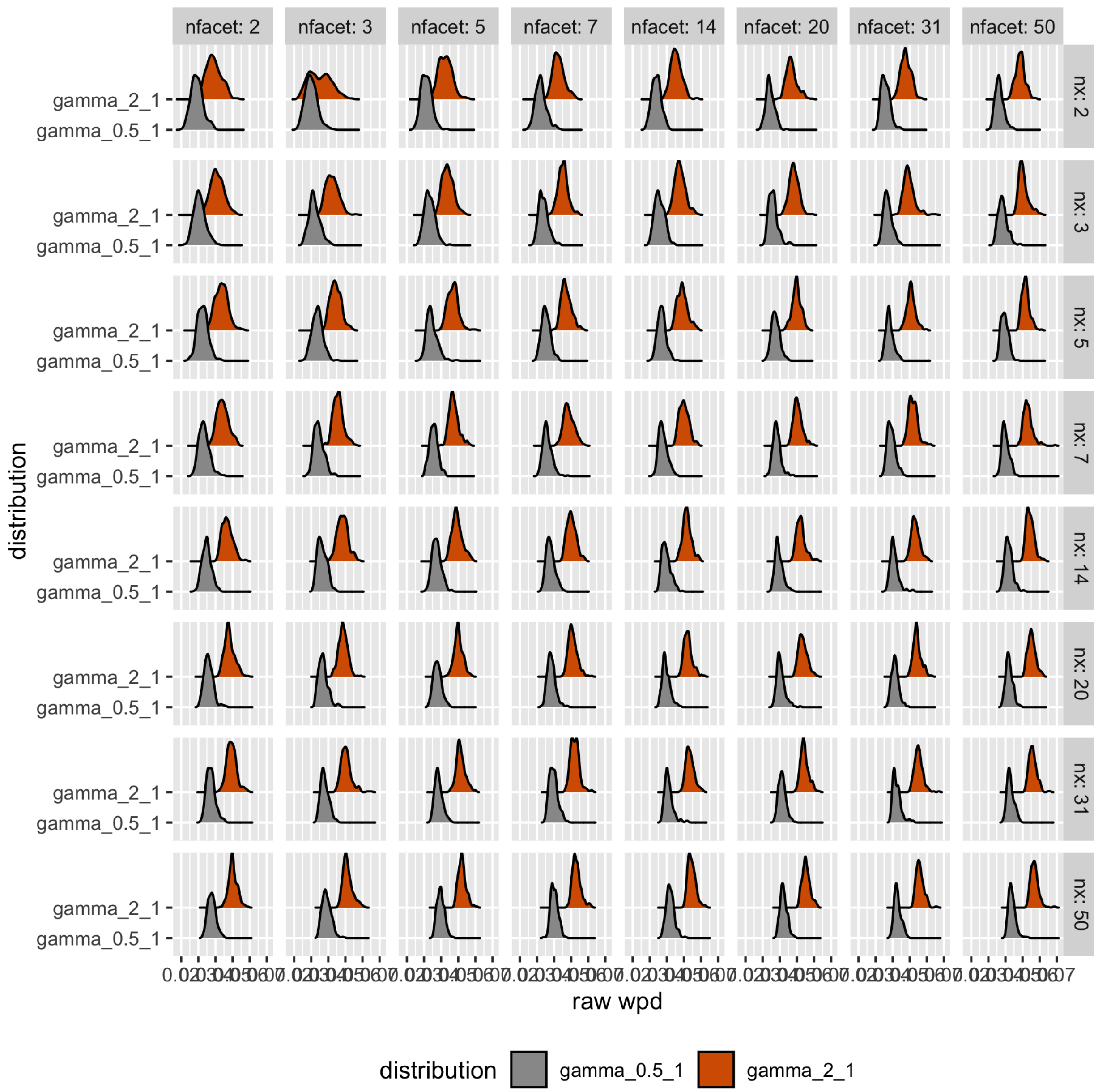


Figure 1: Ridge plots of raw wpd is shown for Gamma(0.5,1), Gamma(2,1) distribution without NQT. The densities change across different facet and x levels and also looks different for the two distributions, which implies wpd value is affected by the change in the shape paramter of the gamma distribution.

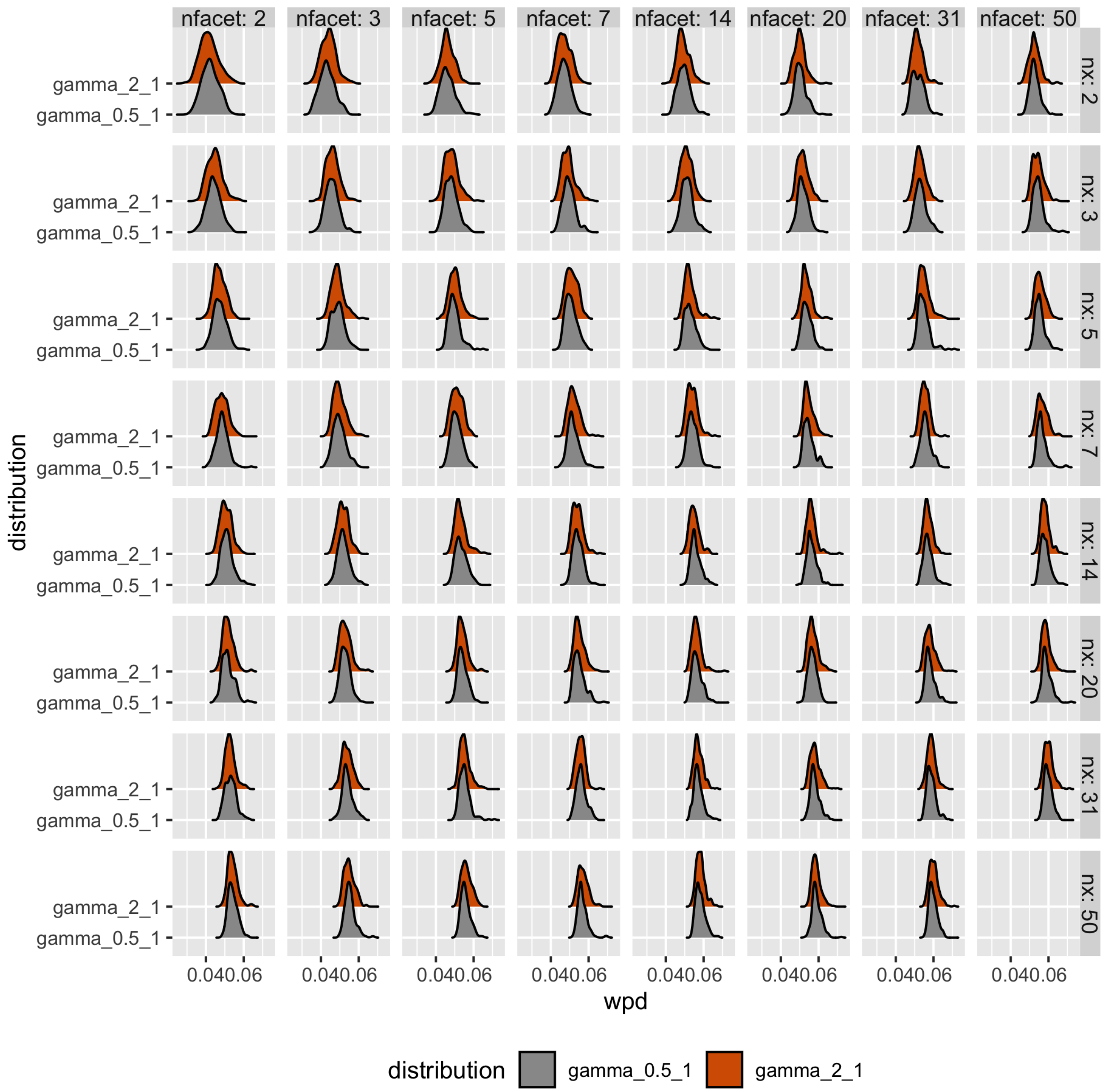


Figure 2: Ridge plots of raw wpd is shown for Gamma(0.5,1), Gamma(2,1) distribution. The densities change across different facet and x levels but look same for the two distributions, which implies wpd value is unaffected by the change in the shape paramter of the gamma distribution

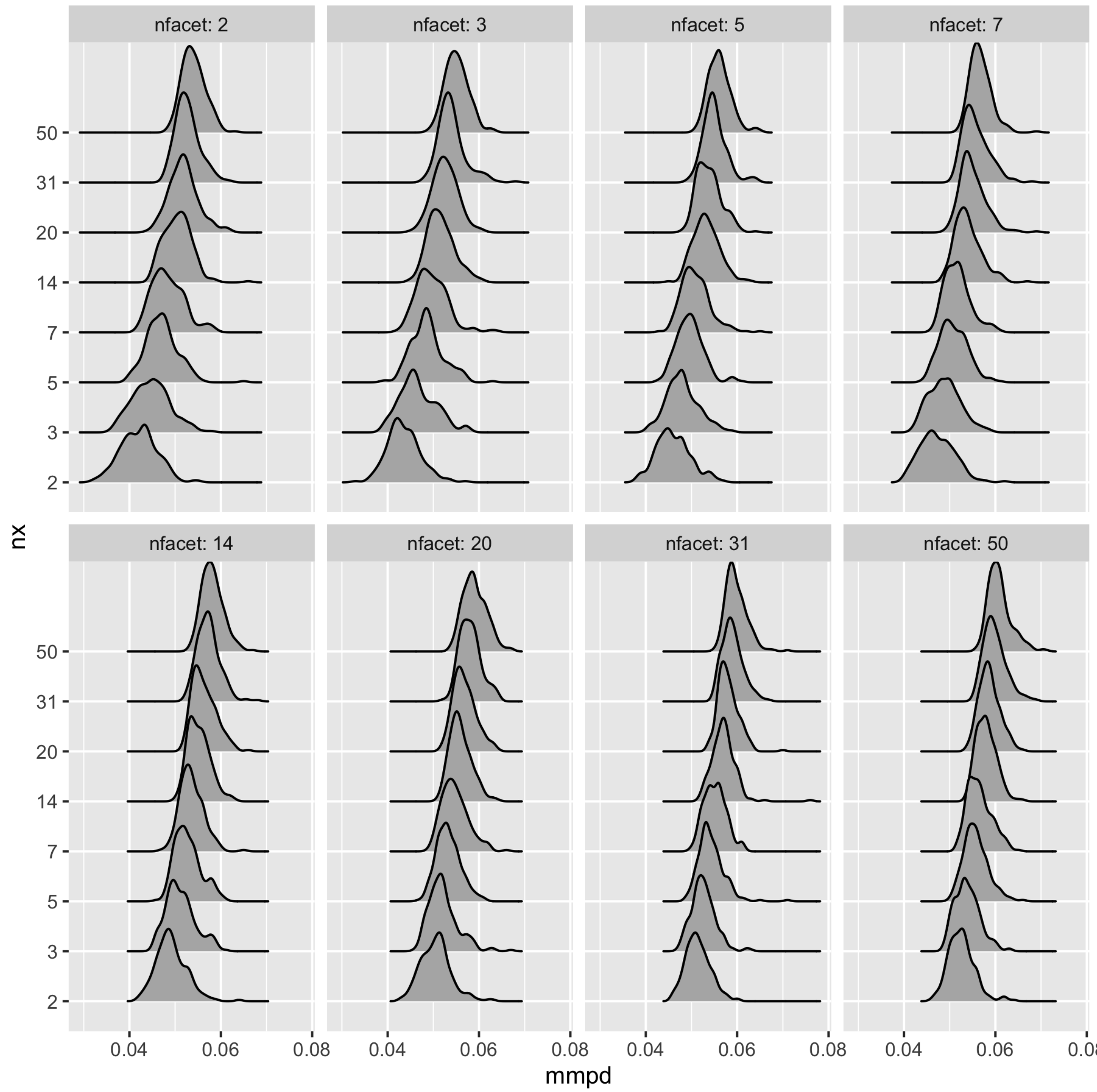
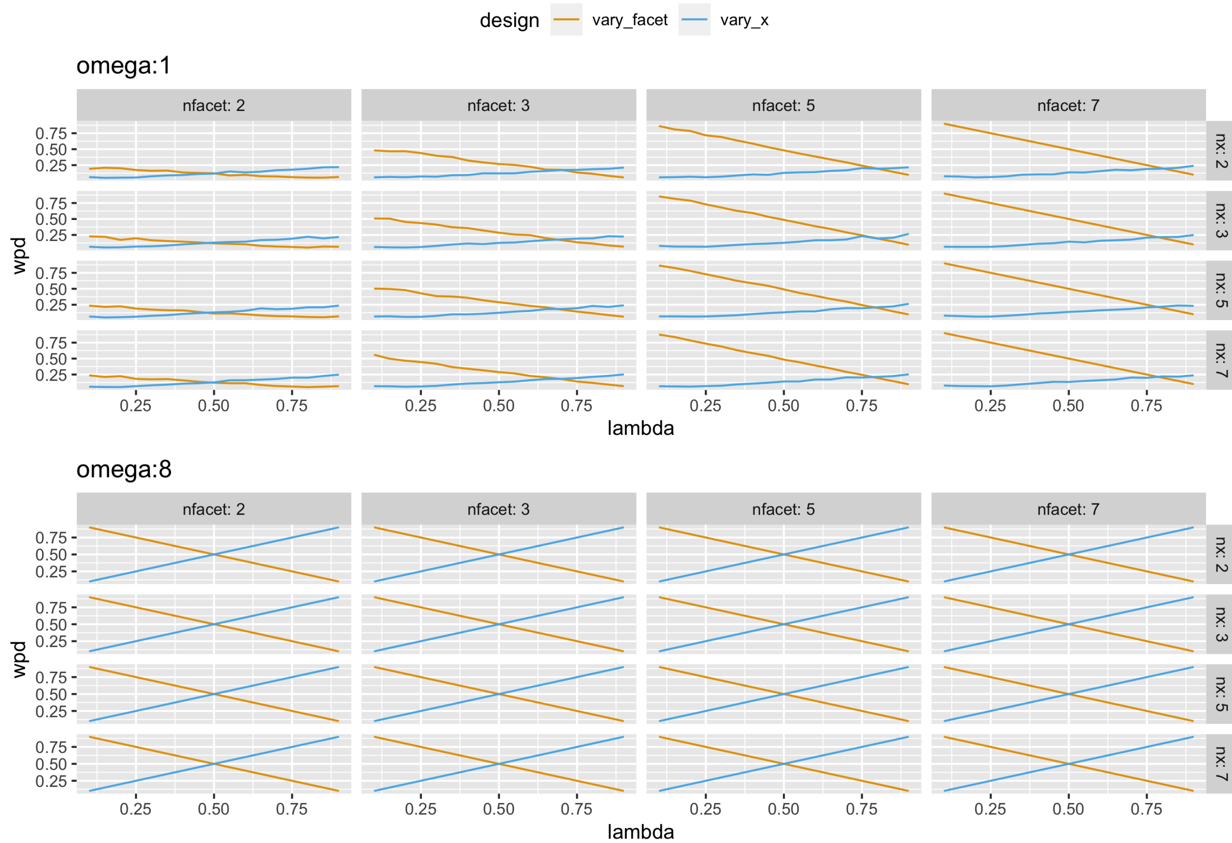


Figure 3: Ridge plots of raw wpd is shown for N(0,5) distribution. For each panel, it could be seen that the location shifts to the right for increasing x levels. Across each panel, the scale of the distribution seems to change for low/moderately lower values and higher values of nfacets and left tails are longer for lower facet levels.

3.4 Tuning parameter



3.5 Increment

4 Behavior of normalised distance measure

4.1 sample size

4.2 number of permutations

4.3 designs

5 Ranking and selecting harmonies

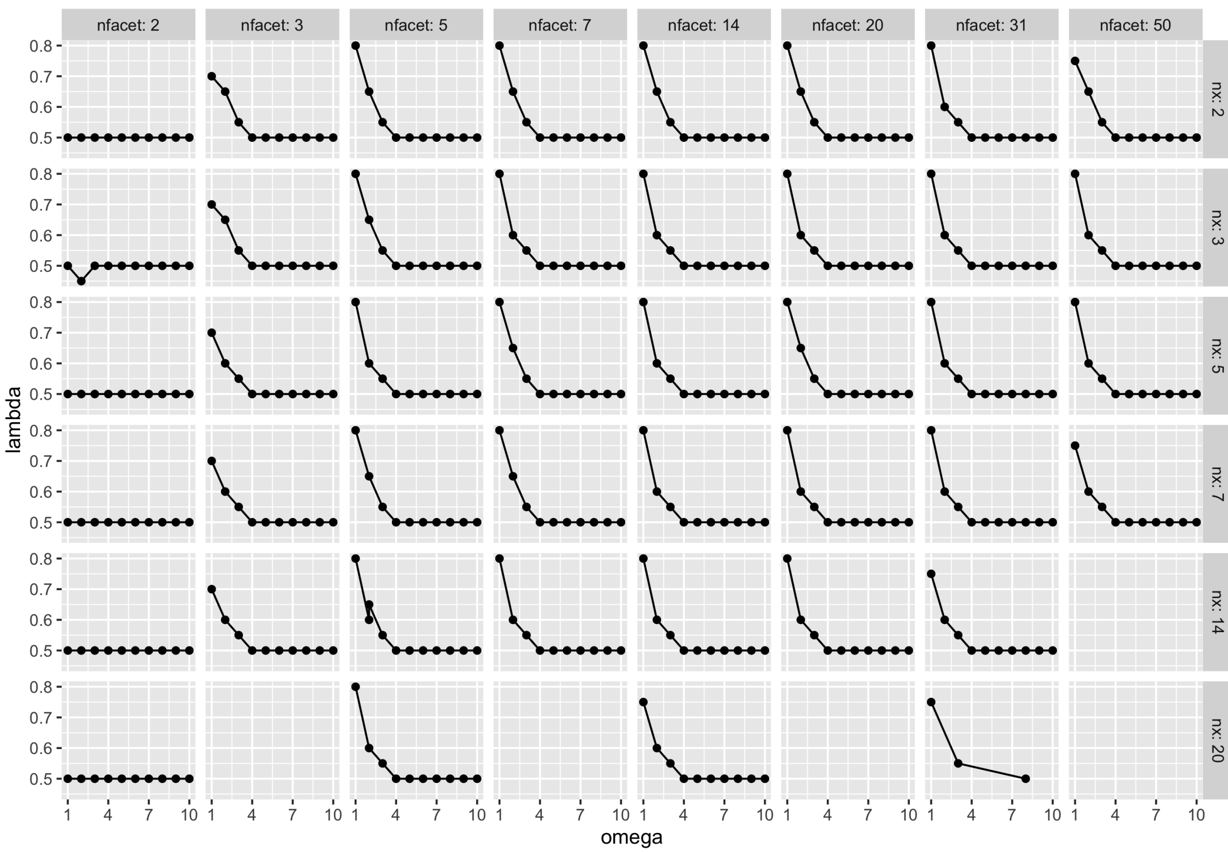


Figure 4: For most panels it is observed that the most common value of the tuning paramter for which the designs interact is 0.5, which implies any value greater than 0.5 could be chosen to up-weigh the within-facet distances and down-weigh the between-facet distances for most sitatuions.