Combining the permutation and modeling approaches for normalisation

1 Data generation

Observations are generated from a Gamma(2,1) 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 $\{nx = 2, nfacet = 2\}, \{nx = 2, nfacet = 3\}, \{nx = 2, nfacet = 5\}, \ldots, \{nx = 50, nfacet = 31\}, \{nx = 50, nfacet = 50\}$. 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 λ is set to 0.67 and values of raw wpd wpd_{raw} is obtained.

Figure 1 shows the distribution of wpd_{raw} plotted across different nx and nfacet categories. Both shape and scale of the distributions change across panels. This is not desirable as it would mean we would not be able to compare wpd_{raw} across different nx and $nfacet$ as each of them are drawn from distributions with different locations and scale. In Figure 2, we see how the median of wpd_{raw} varies with the total number of distances $nx * nfacet$ for each panel. The median increases abruptly for lower values of $nx * nfacet$ and slowly for higher $nx * nfacet$.

We need a transformation on raw_{wpd} which will make it independent of the values of $nx * nfacet$. Two approaches have been employed for that purpose, the first one involves fitting a model and the latter involves a permutation method to make the distribution of the transformed wpd_raw similar across different nx and $nfacet$.

1.1 Modeling approach to normalisation

1.1.1 Linear model

A linear model is fitted to see how the values of wpd_{raw} changes with the values of nx and $nfacet$. The model is of the form

$$y = a + b * log(x) + e$$

, where $y = median(wpd_{raw})$ and $x = nx * nfacet$. wpd_{lm} is a transformation on wpd_{raw} which is designed to remove the impact of $nx * nfacet$ on wpd_{raw} and thus is defined as follows: $wpd_{lm} = wpd_{raw} - \hat{a} - \hat{b} * log(nx * nfacet)$ $wpd - lm - horizontal$ seems to have no relationship with $nx * nfacet$ as could be seen in Figure 3.

1.1.2 Generalised linear model

In the linear model approach, $wpd_{raw} \in R$ was assumed, whereas, wpd_{raw}, Jensen-Shannon Distance (JSD) lies between 0 and 1. Furthermore, JSD follows a Chi-square distribution, which is a special case of Gamma distribution and hence belongs to exponential family of distributions. Therefore, we can fit a generalized linear model instead of a linear model to allow for the response variable to follow a Gamma distribution.

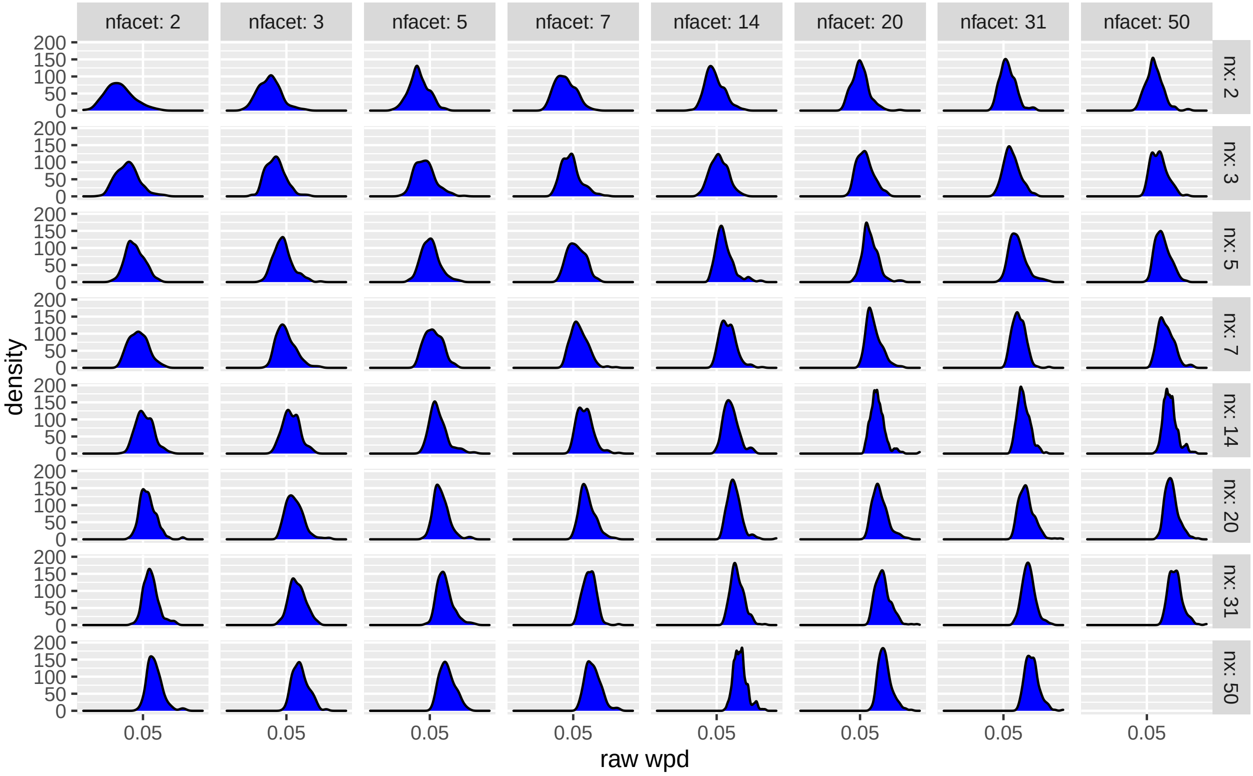


Figure 1: Distribution of raw wpd is plotted across different nx and nfacet categories. Both shape and scale of the distribution changes for different nx and nfacet categories.

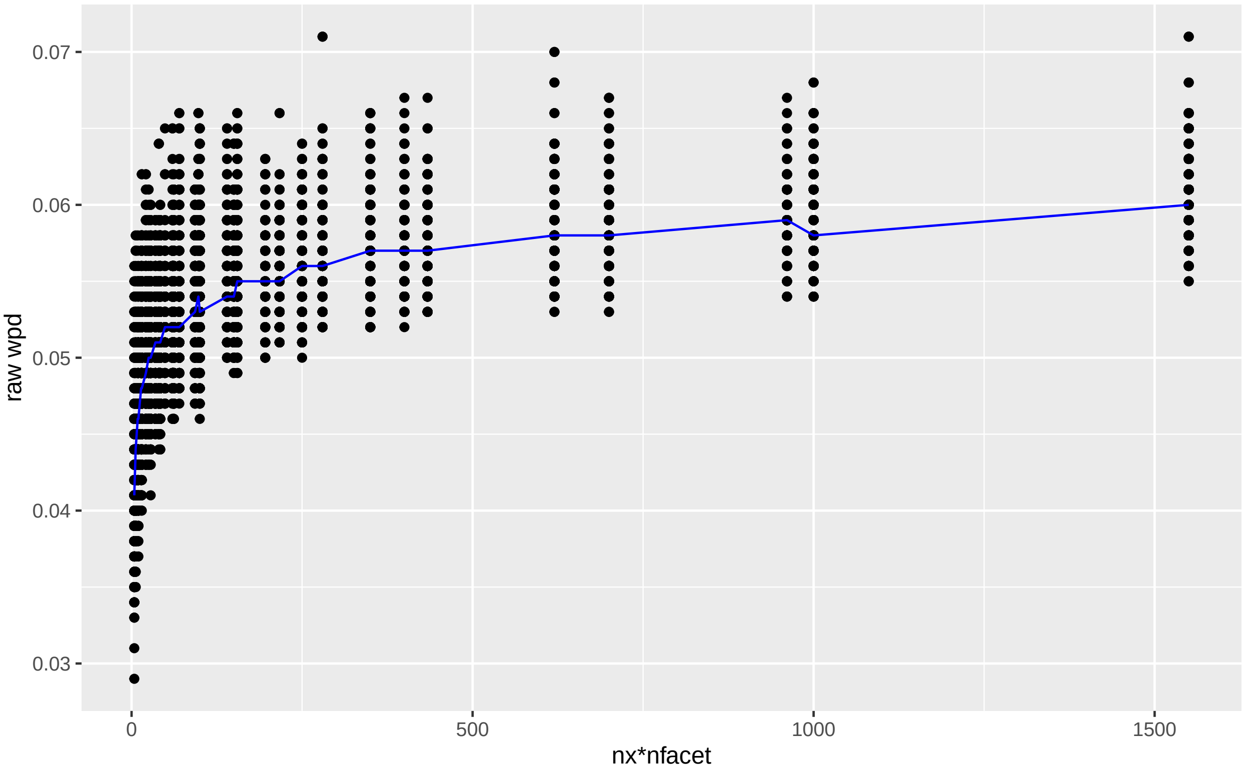


Figure 2: wpd_{raw} is plotted against nx*nfacet and the blue line represents the median of the multiple values for each nx*nfacet levels.

The inverse link is used when we know that the mean response is bounded, which is applicable in our case since $0 \leq wpd_{raw} \leq 1$.

We fit a Gamma generalized linear model with the inverse link which is of the form:

$$y = a + b * log(x) + e$$

, where $y = median(wpd_{raw})$, $x = nx * nfacet$. Let $E(y) = \mu$ and $a + b * log(x) = g(\mu)$ where g is the link function. Then $g(\mu) = 1/mu$ and $\hat{\mu} = 1/(\hat{a} + \hat{b}log(x))$. The residuals from this model $(y - \hat{y}) = (y - 1/(\hat{a} + \hat{b}log(x)))$ would be expected to have no dependency on x. Thus, wpd_{glm} is chosen as the residuals from this model and is defined as: $wpd_{glm} = wpd_{raw} - 1/(\hat{a} + \hat{b} * log(nx * nfacet))$.

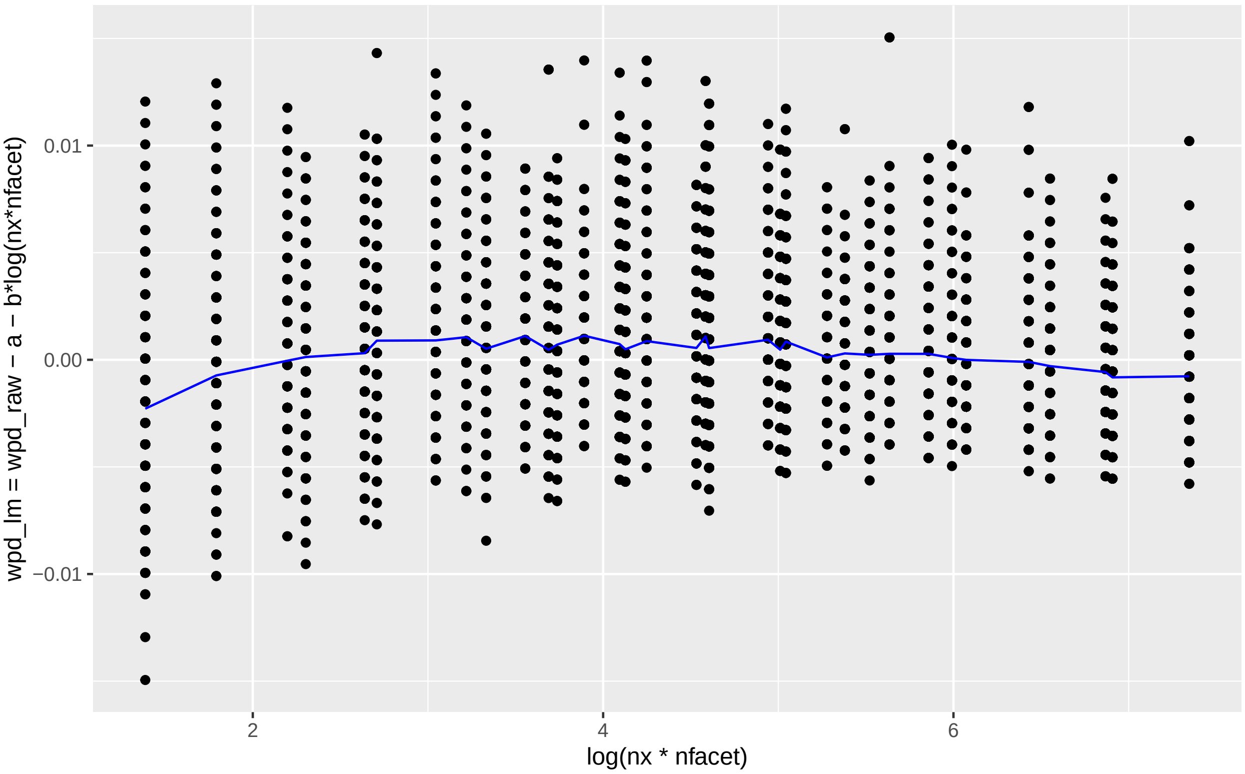


Figure 3: wpd_{lm} is plotted against log(nx*nfacet) and this transformation leads to median(wpd_{lm}) to having almost no relationship with log(nx*nfacet)

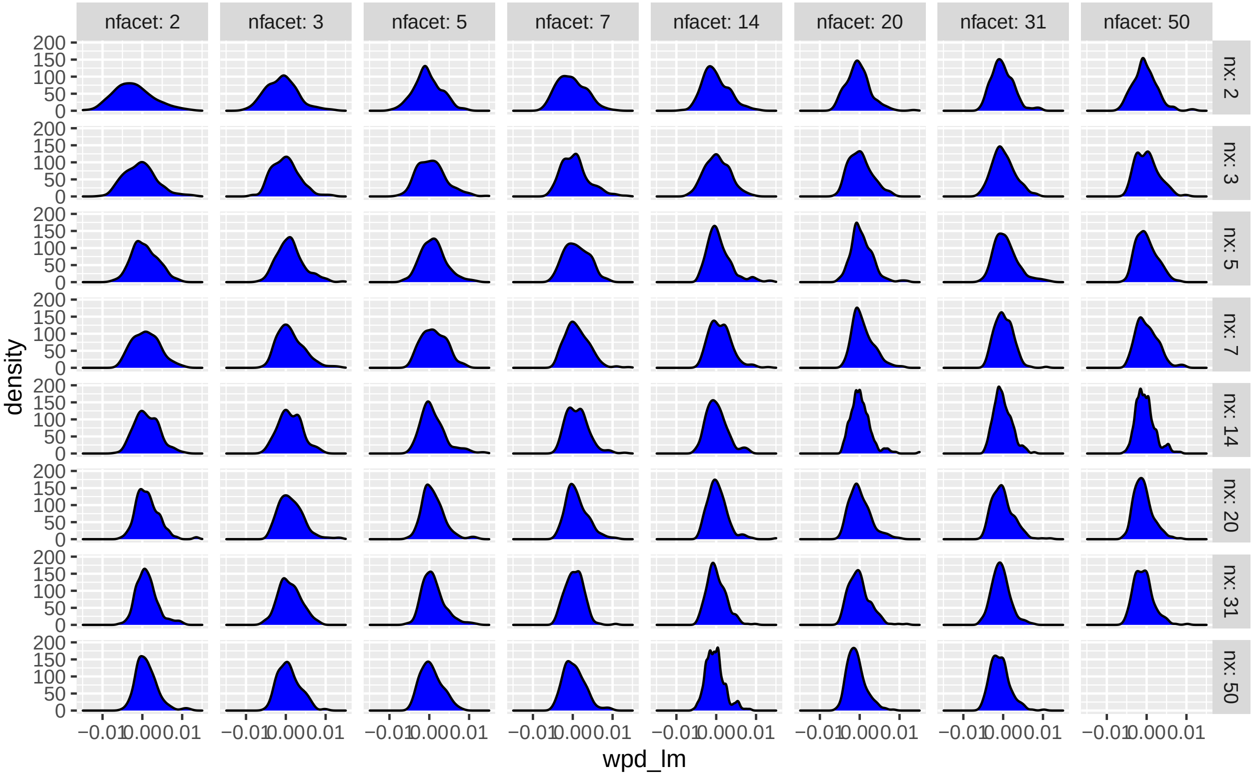
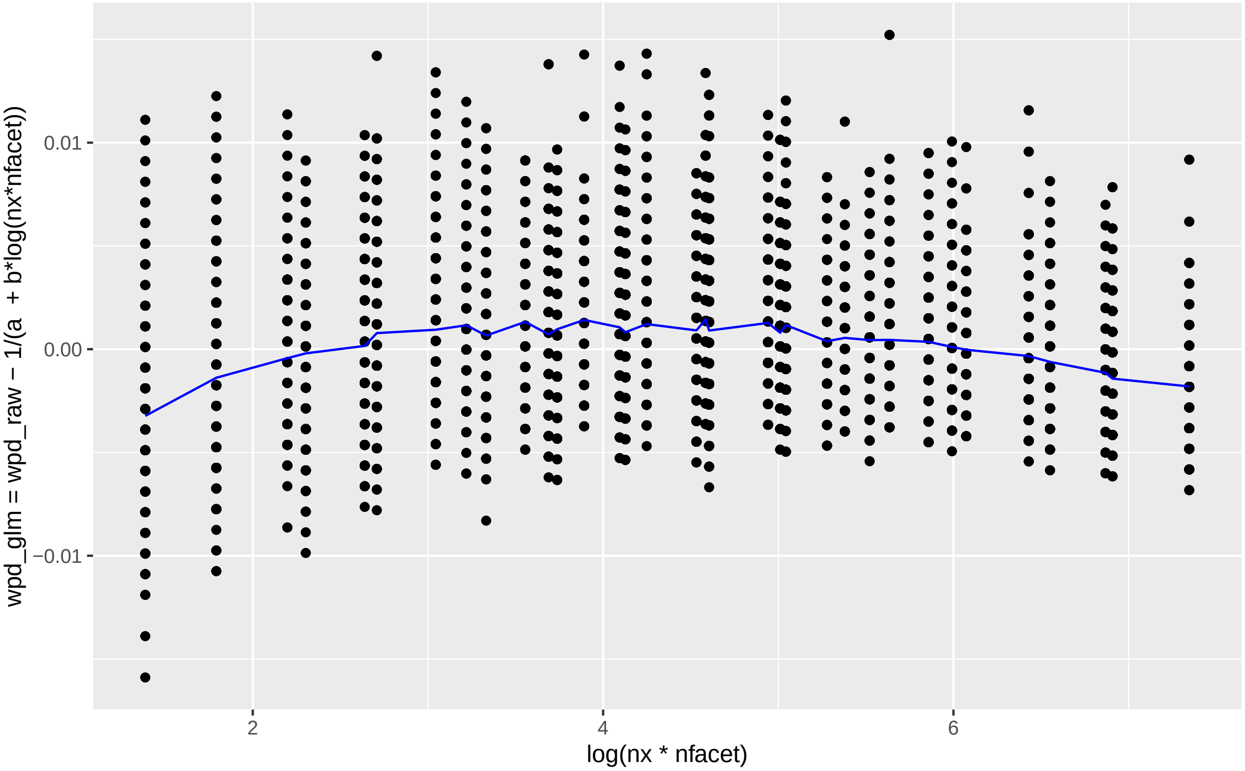


Figure 4: The distribution of wpd_{lm} is plotted. The distributions are more similar across higher nx and nfacet and are different for smaller nx and nfacet.



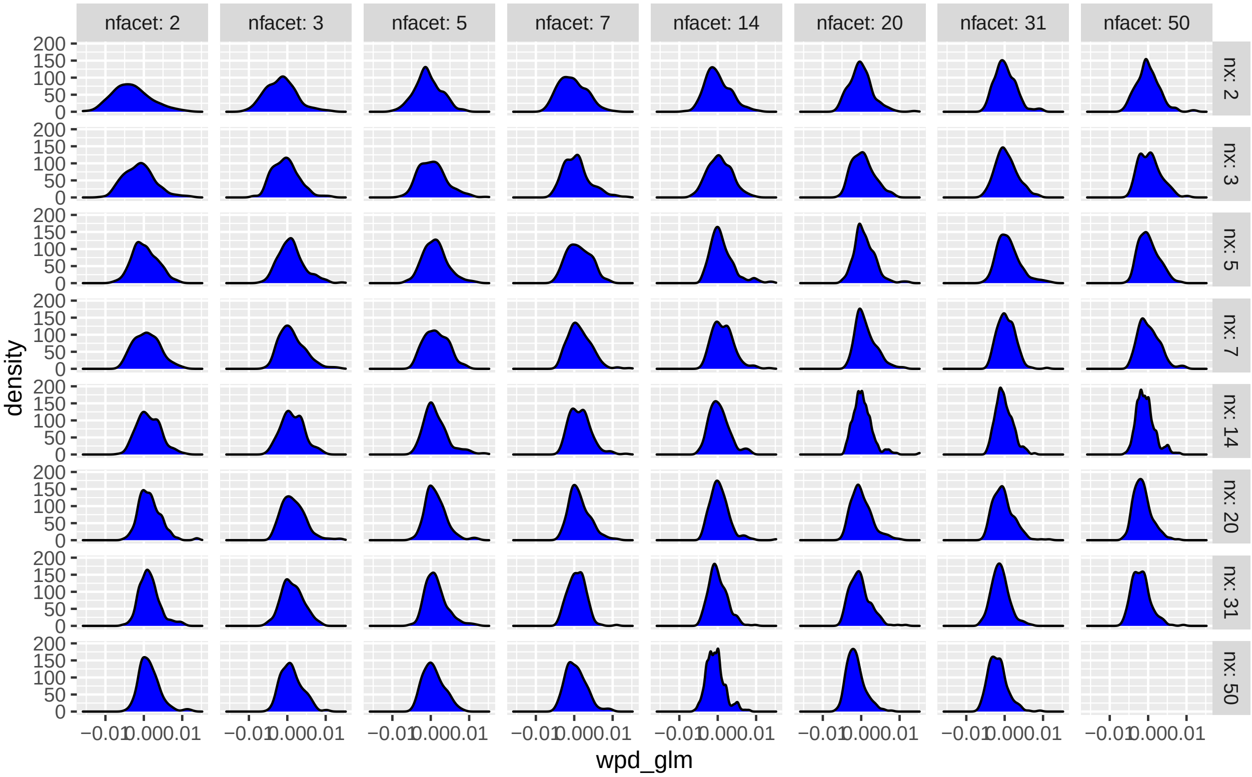


Figure 5: The distribution of wpd_{glm} is plotted. The distributions are more similar across higher nx and nfacet and dissimilar for fewer nc and nfacets.

1.2 Permutation approach to normalisation

The simulated data for each of the panels is permuted/shuffled $nperm = 200$ times and for each of those permutations wpd_{norm} is computed as follows: $wpd_{perm} = (wpd_{raw} - mean(wpd_{raw}))/sd(wpd_{raw})$. This is done so that the distribution of the normalised measure wpd_{norm} has the same mean and standard deviation across different nx and nfacet.

Please note that standardizing the variable wpd_{perm} in this approach leads to $location = 0$ and $scale = 1$ for this variable.

1.3 Bringing them both to the same scale

We see that the transformation through the modeling approach leads to very similar distribution across high nx and $nfacet$ (higher than 7) and not so much for lower nx and $nfacet$. Hence, the computational load of permutation approach could be alleviated by using the modeling approach for the higher nx and $nfacet$, however, it is important that we use the permutation approach for lower nx and $nfacet$. However, it is difficult to compare the transformed wpd from both of these approaches, since each of the variables is measured on a different scale (each of them have location 0). The transformed variables from the two approaches could be brought to the same scale so that for smaller categories, permutation approach is used and for larger categories, we can stick to modeling approach. These could be done through the following:

- Making the range of both the variables same by using min-max scaling method. In practice, however, we would only have one value of wpd_{raw} which we need to transform using the modeling approach. Hence, min-max scaling approach could not be used here.

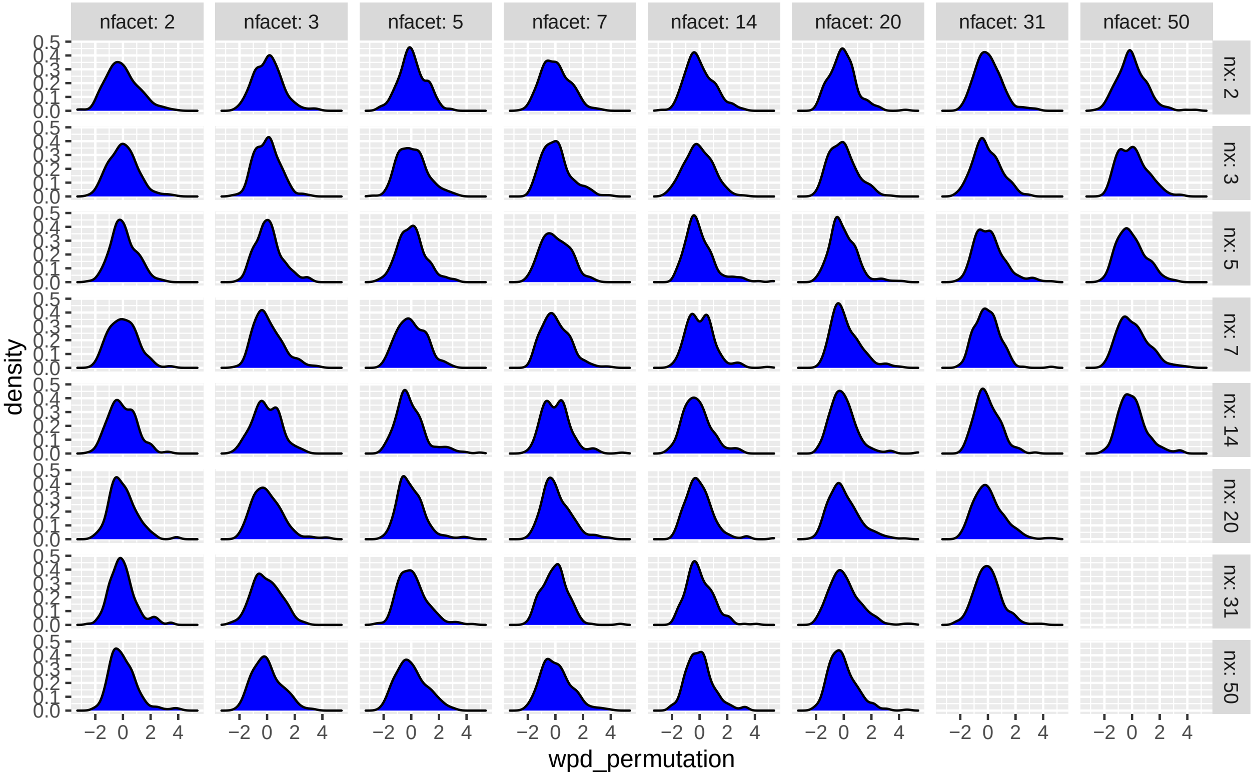


Figure 6: The distribution of $wpd_{permutation}$ is plotted. The distributions are more similar across different nx and nfacet (specially for small nx and nfacet) but this approach has the downside of more computational time.

- Standardizing the variables and expressing scores at standard deviation units. Again in practice, however, we would only have one value of wpd_{raw} which we need to transform using the modeling approach. Hence, standardizing scores could not be used here as we do not have the mean and standard deviation of a series while using transformation using modeling.
- Make the location and scale of both the approaches similar so that they could be compared. Please note that the range of values could be different in this case, however location and scale are brought to same levels.)

The measure wpd_{glm} has location 0 and standard deviation ~ 0.003, whereas the measure $wpd_{permutation}$ which is a z-score, has a normal distribution with location 0 and standard deviation 1. To bring them to the same scale, we have defined $wpd_{glm-scaled} = wpd_{glm} * 300$, which brings the standard deviation of $wpd_{glm-scaled}$ to almost 1, without changing the location.

The measure $wpd_{glm-scaled}$ seems to roughly follow a normal distribution except in the tails as could be seen in Figure 7 and the very method of permutation approach ensures that $wpd_permutation$ is also normally distributed. Further, they are brought to the similar scale and location and hence could be compared.

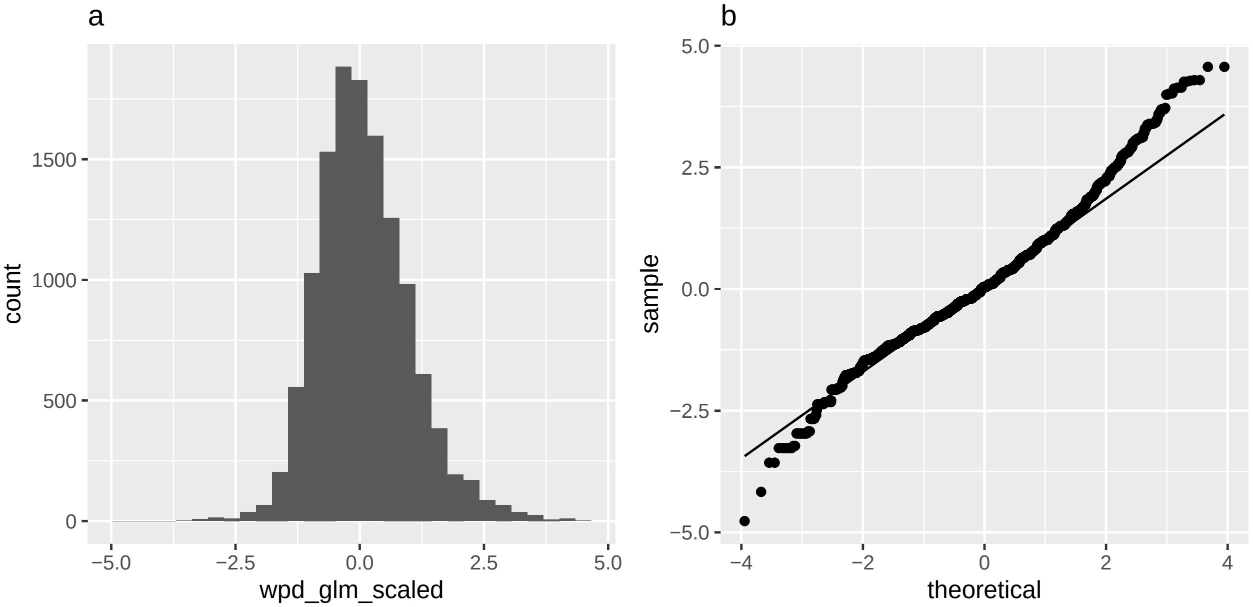


Figure 7: In panel a, the histogram of $wpd_{glm-scaled}$ is plotted. In parl b, the QQ plot is shown with the theoretical quantiles on the x-axis and $wpd_{glm-scaled}$ quantiles on the y-axis. The distribution looks symmetric and looks like normal except in the tails.

Questions:

1. Does the approach of wpd_{glm} looks correct?
2. Are wpd_{glm} and $wpd_{permuation}$ be compared from this plot or both should be brought to the scale 1 for comparison?
3. Forcing same range (0,1) to both wpd_{glm} and $wpd_{permutation}$ could be obtained by using the transformation $(z - z_{min})/(z_{max} - z_{min})$. This could be used for the permutation approach. But the modeling approach would only have one value of wpd_{raw} for a panel in practice, how to scale the values in the modeling approach so that they are within 0 and 1? Getting them to 0 and 1 is changing the location making the comparison difficult.

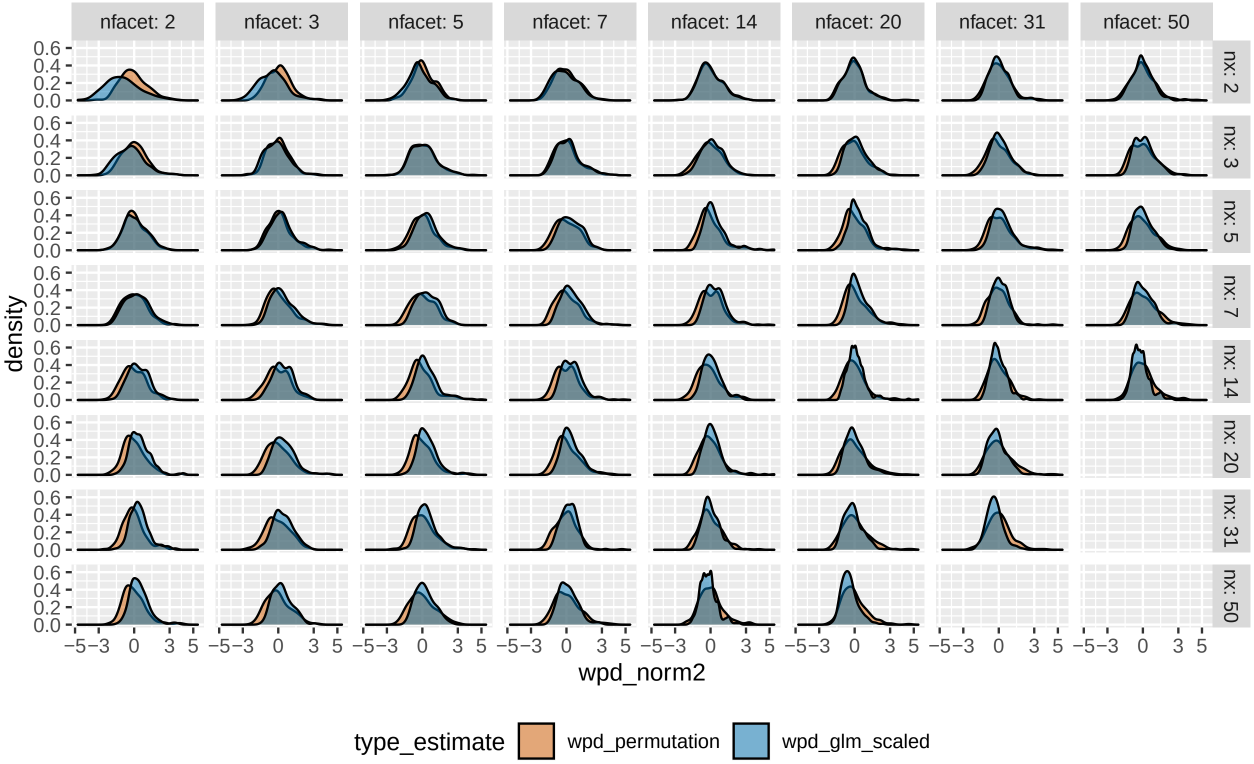


Figure 8: w$pd_{p}ermutation$ and $wpd_{glm-scaled}$ are plotted together on the same scale. They also have the same location and hence the values from these two approaches could be compared across panels. $wpd_{glm-scaled}$ would be used to normalise wpd_{raw} for higher nx and $nfacet$ and $wpd_{glm-scaled}$ would be used for smaller levels to alleviate the problem of computational time.