Compare permutation and scalar transformation approaches on simulated and real data

Sayani Gupta

1 Simulated data

1.1 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.

1.2 Scalar transformation approach to normalisation

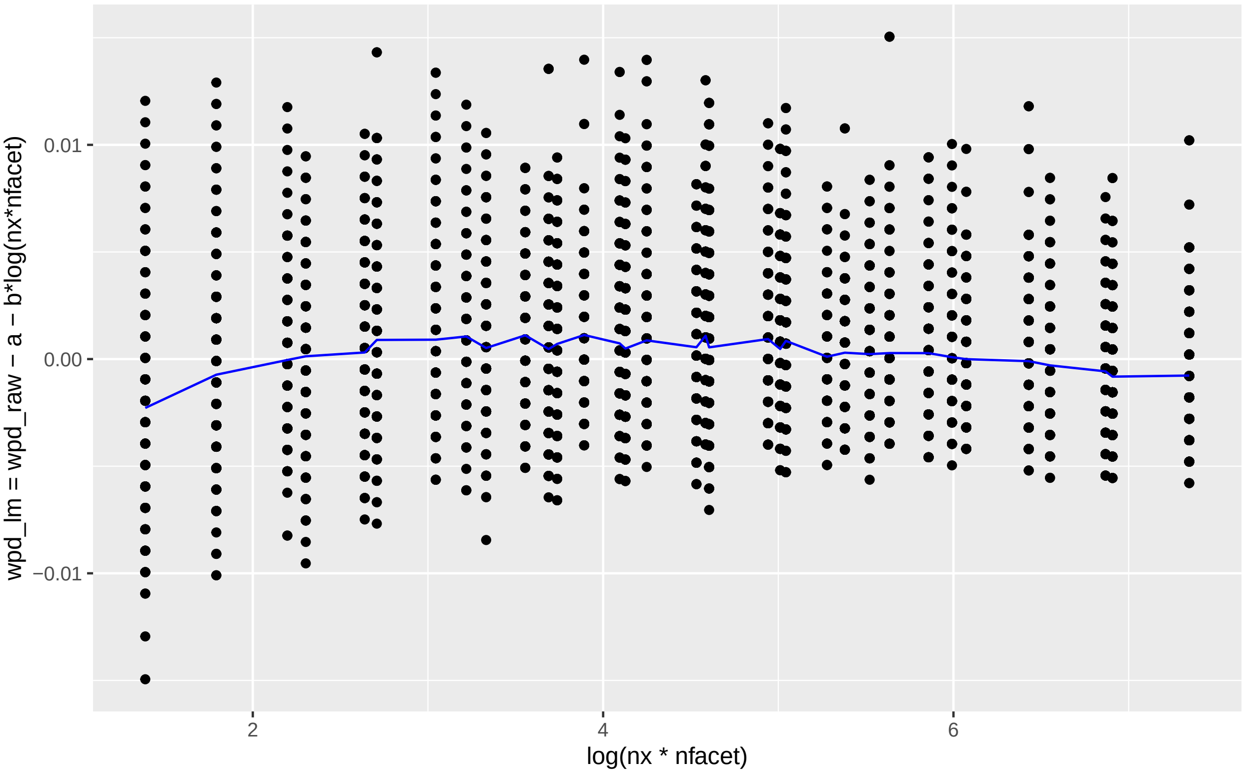
1.2.1 Linear model

A log-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 should be designed to remove the effect of $nx * nfacet$ on wpd_{raw} and thus is defined as follows: $wpd_{lm} = wpd_{raw} - a - b * log(nx * nfacet)$

```
##
## Call:
## lm(formula = actual ~ poly(log(`nx * nfacet`), 1, raw = TRUE),
##     data = G21_median)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -2.946e-03 -2.240e-04  5.135e-05  4.147e-04  1.014e-03
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             4.003e-02  4.171e-04   95.97   <2e-16
## poly(log(`nx * nfacet`), 1, raw = TRUE) 2.826e-03  8.796e-05   32.13   <2e-16
##
## (Intercept)                             ***
## poly(log(`nx * nfacet`), 1, raw = TRUE) ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0007881 on 32 degrees of freedom
## Multiple R-squared:  0.9699, Adjusted R-squared:  0.969
## F-statistic:  1032 on 1 and 32 DF,  p-value: < 2.2e-16
```



1.2.2 Generalised linear model

In the earlier approach, $wpd_{raw} \in R$, whereas, $0 \leq wpd_{raw} \leq 1$ since it is a JS distance. Also, JSD follows a Chi-square distribution, which could be considered as a Gamma distribution. Hence, to restrict the range of the measure between 0 and 1, and deviate from the normality assumption of the response variable in a OLS approach, we fit a generalized linear model (GLM) with the error distributed as a gamma distribution and link function as "inverse". Hence, the model is of the form

$$1/y = a + b * log(x) + e$$

, where $y = median(wpd_{raw})$ and $x = nx * nfacet$. Again, wpd_{glm} is a transformation on wpd_{raw} which should be designed to remove the effect of $nx * nfacet$ on wpd_{raw} and thus is defined as follows: $wpd_{glm} = 1/wpd_{raw} - a - b * log(nx * nfacet)$.

Please note:

- subtracting the intercept brings the location of the transformed variable to 0 for either case
- heterogeneity is higher for this case after normalization as compared to the earlier one.

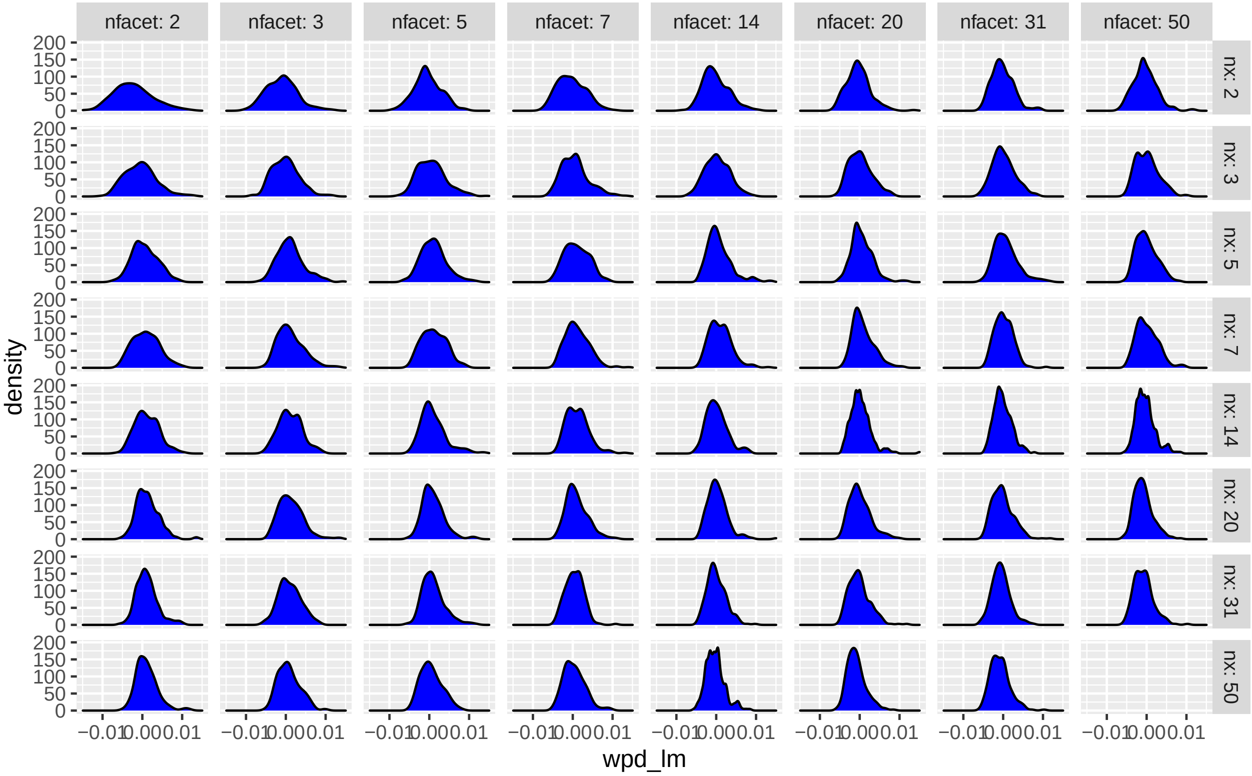
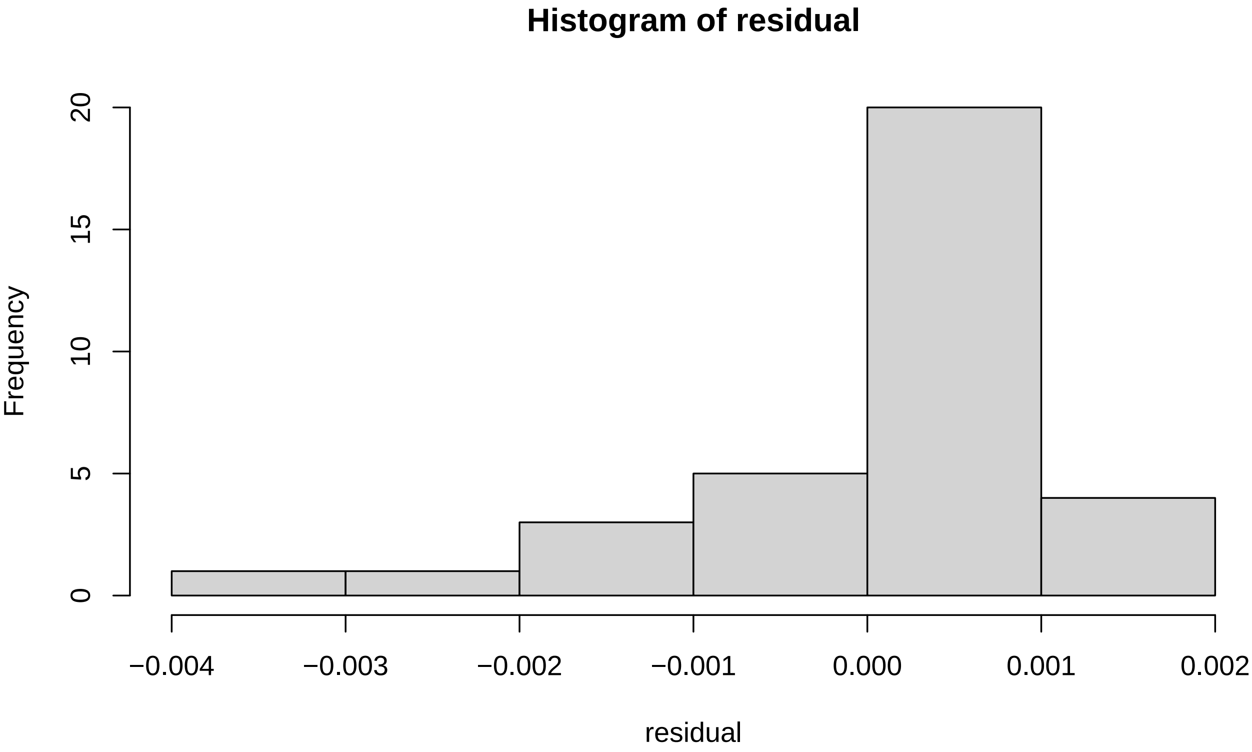
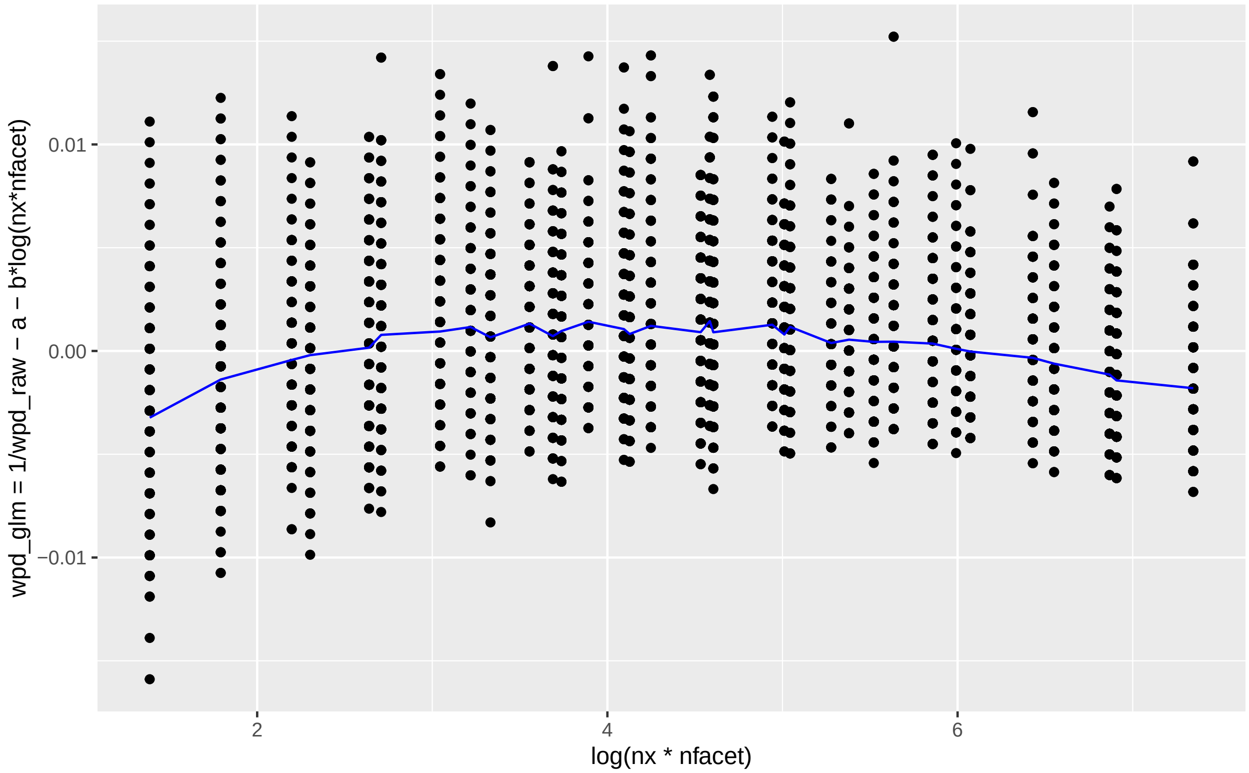
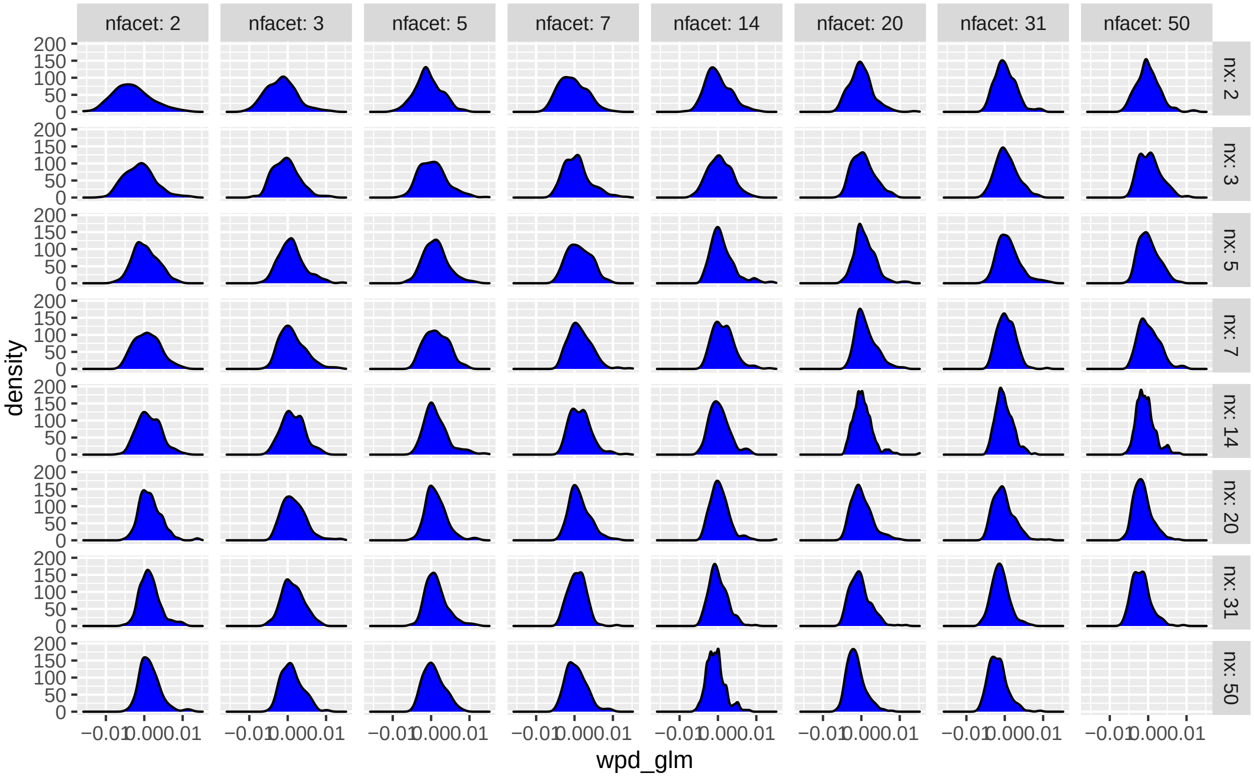
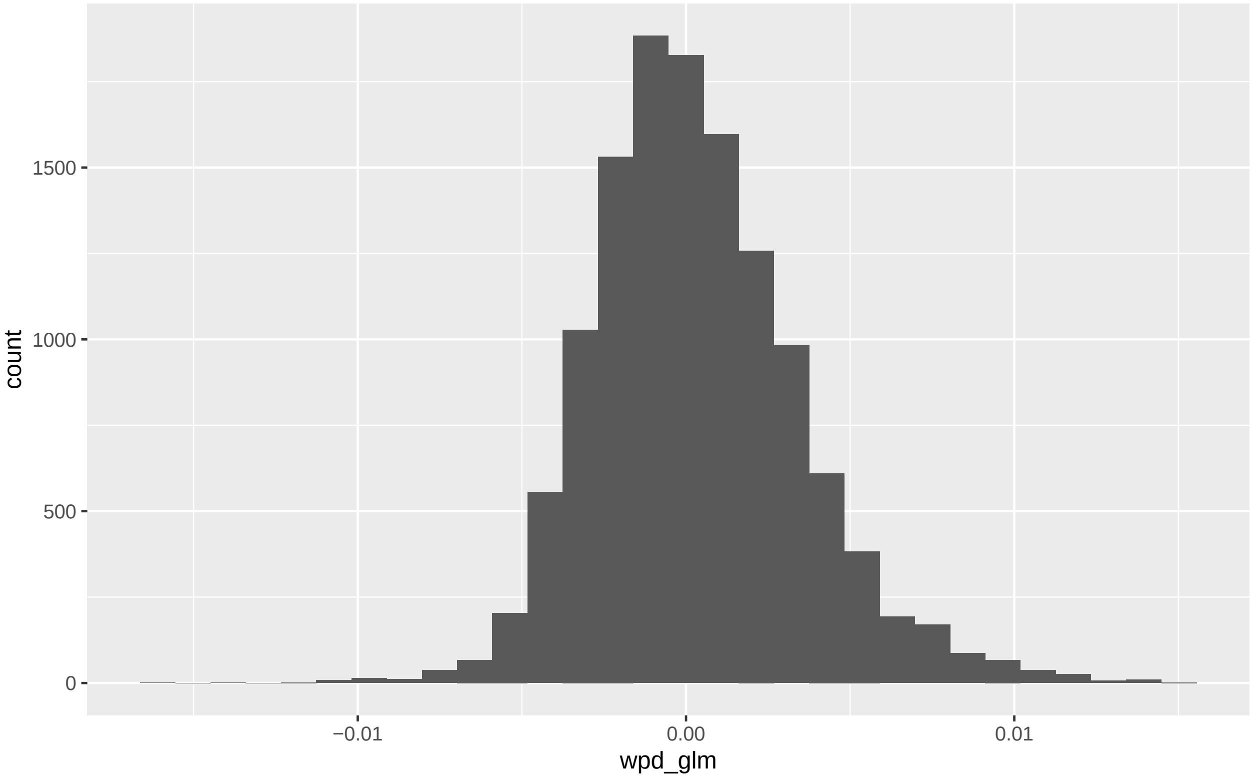


Figure 1: The distribution of wpd_{lm} is plotted. The distributions are more similar across higher nx and nfacet and are different for smaller nx and nfacet.




wpd_glm = 1/wpd_raw − a − b*log(nx*nfacet)
0.01
0.00
−0.01
2
4
6
log(nx * nfacet)


nfacet: 2
nfacet: 3
nfacet: 5
nfacet: 7
nfacet: 14
nfacet: 20
nfacet: 31
nfacet: 50
nx: 2
nx: 3
nx: 5
nx: 7
nx: 14
nx: 20
nx: 31
nx: 50
density
200
150
100
50
0
−0.010.000.01
wpd_glm



```
## [1] -4.767707  4.565350
```

1.3 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.

1.4 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:

- Converting each scale to have the same lower and upper levels
- Standardizing the variables and expressing scores at standard deviation units

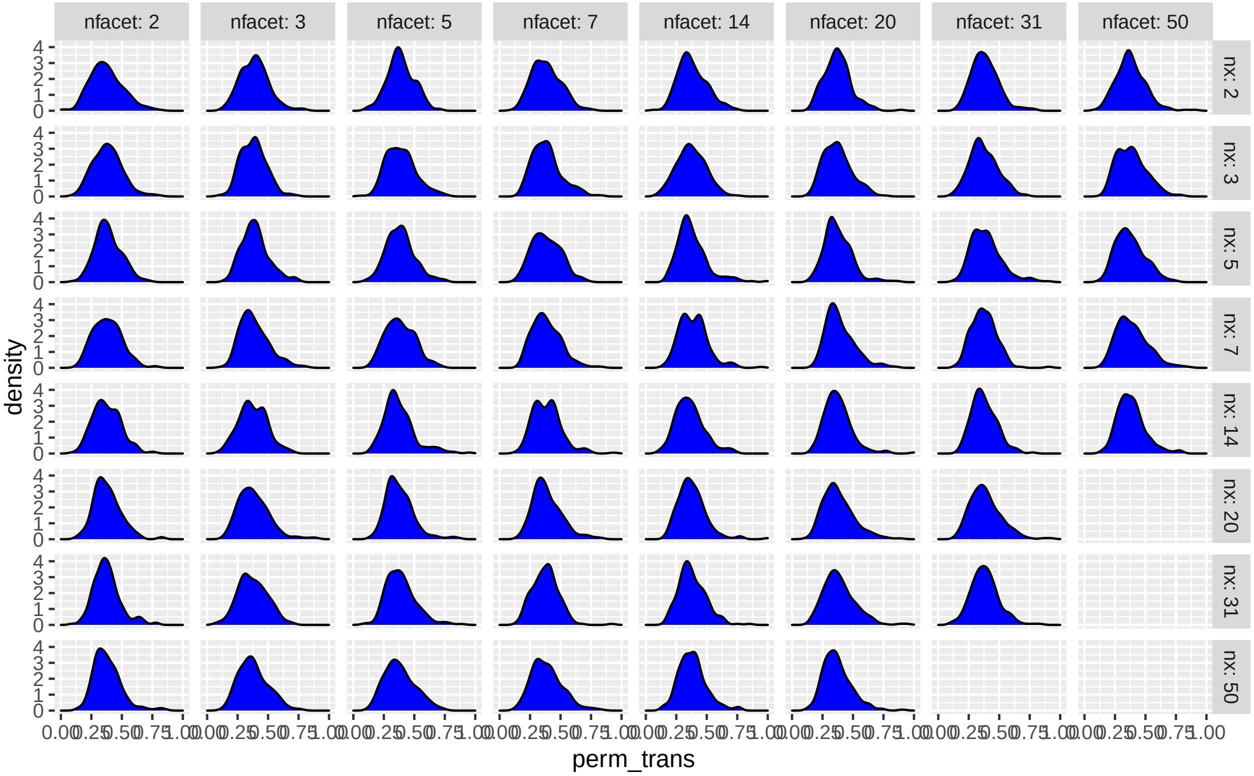


Figure 2: The distribution of $wpd_{permutation}$ is plotted. The distributions are more similar across different nx and nfacet (specially for small nx and nfacet) but this approach has the downside of more computational time.

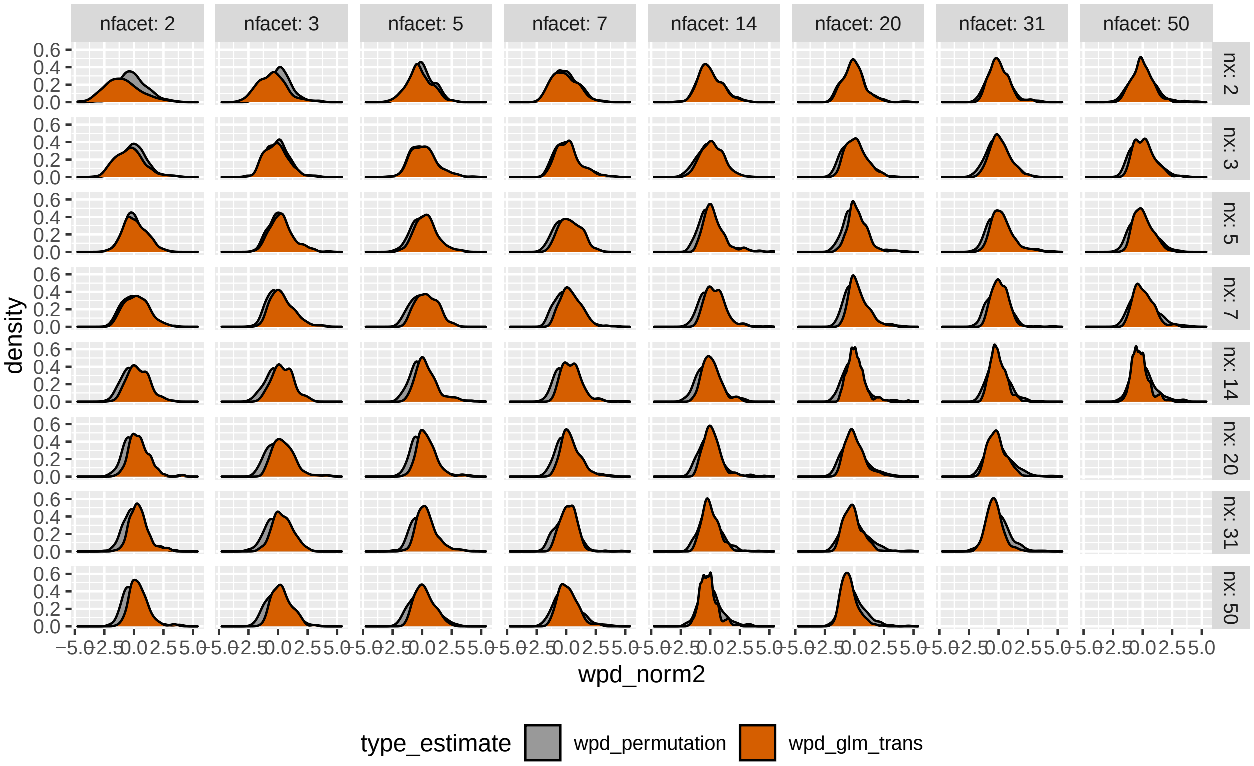


Figure 3: wpd_permutation and wpd_lm are in very different scale and could not be plotted together without transforming further. Whereas wpd_permutation and wpd_glm could be plotted together as they at least have the same location. Are they on the same scale though? Can they be compared?

- making the range of both the variables same (additional advantage that we can have a bound for the normalised measure which could be compared across normalizing approaches and data sets.)

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?

```
## [1] 0.003051935
```