Choosing an appropriate scalar transformation to normalise wpd

Sayani Gupta

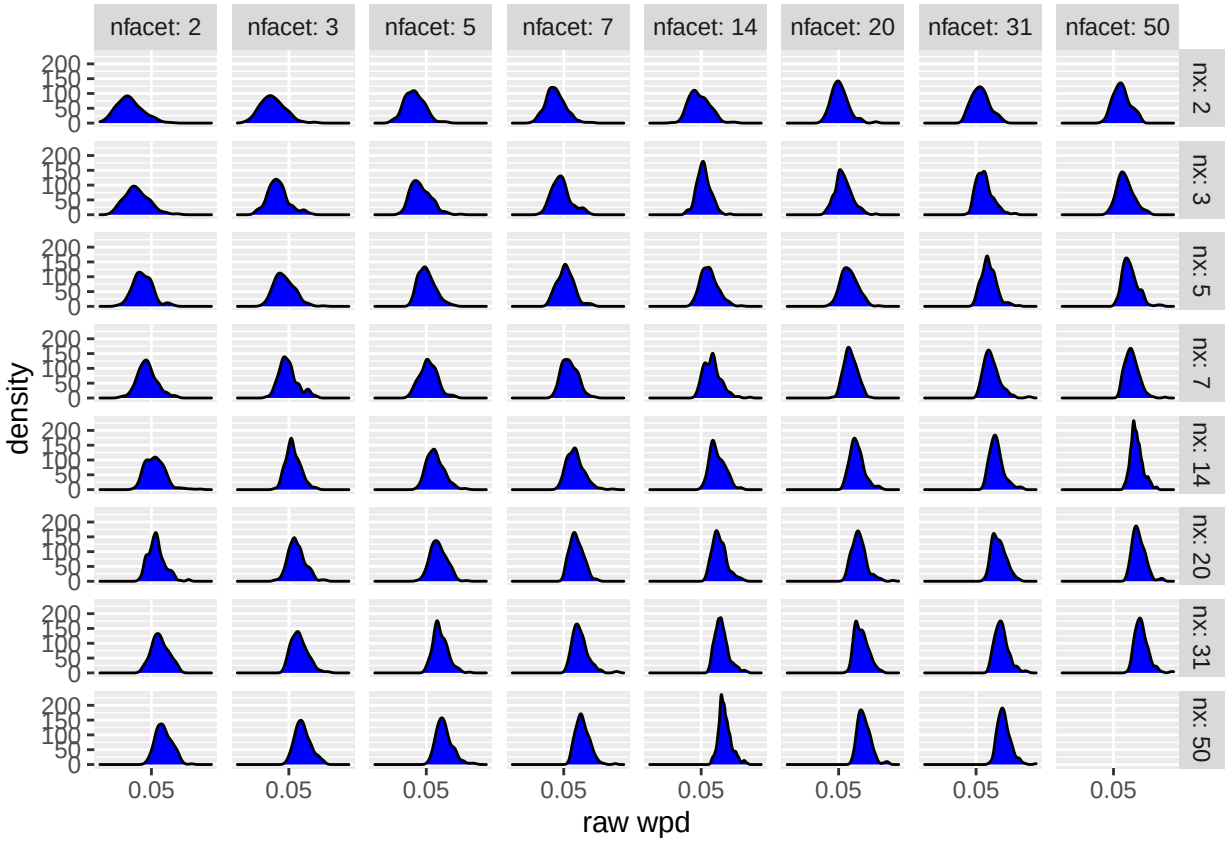
28/01/2021

- Data presented

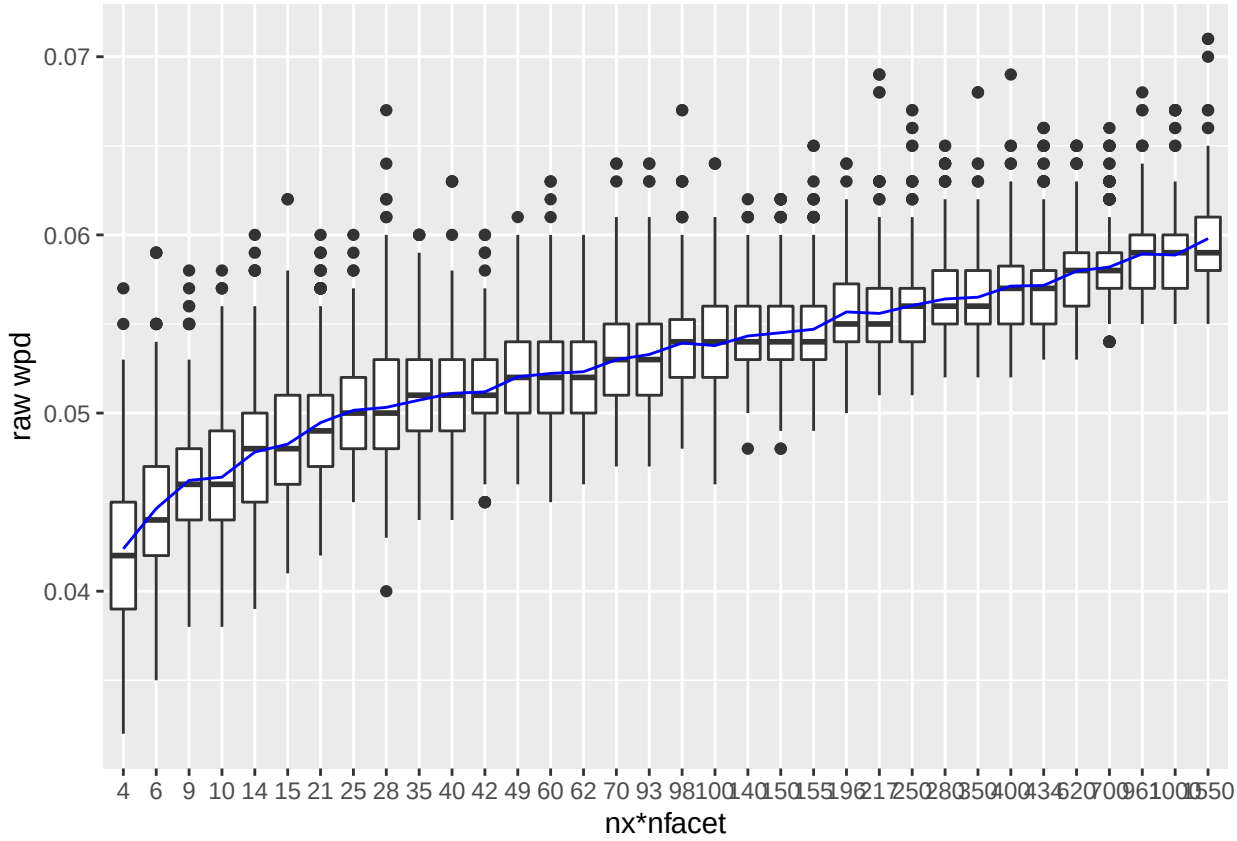
Observations are generated from a N(0,1) distribution for each combination of nx and $nfacet$ from the following sets: $nx = nfacet = \{2, 3, 5, 7, 14, 20, 31, 50\}$ to cover a wide range of levels from very low to moderately high. Each combination is being referred to as a *panel*. That is, data is 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 is obtained.

- How the distribution of the raw wpd looks across nfacets and nx?

Both shape and scale of the distribution changes for different nx and nfacet categories.

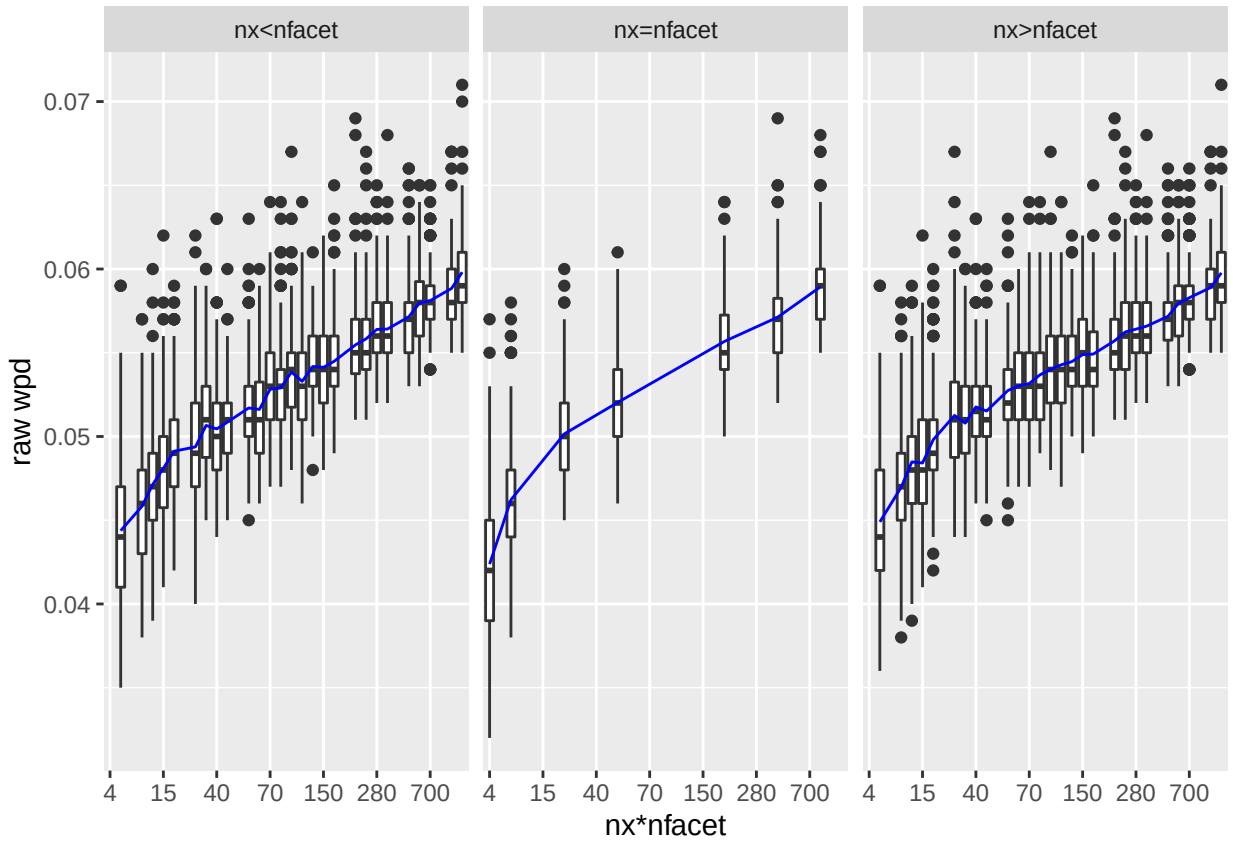


- Plot the values of wpd against nx*nfacet to see the rough relationship



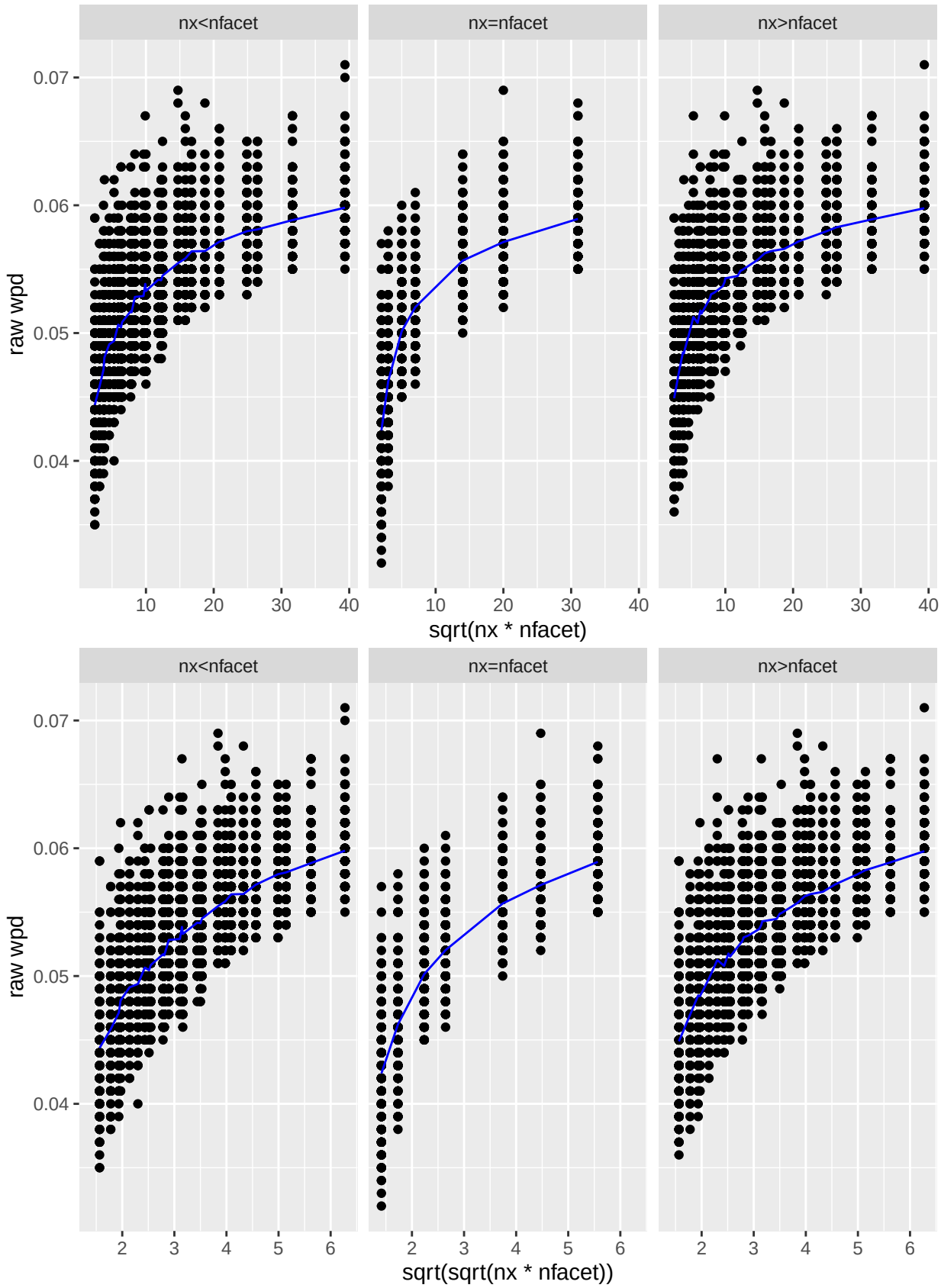
- Plot the values of wpd against nx*nfacet to see if the same relationship holds for different nx and nfacet relationships

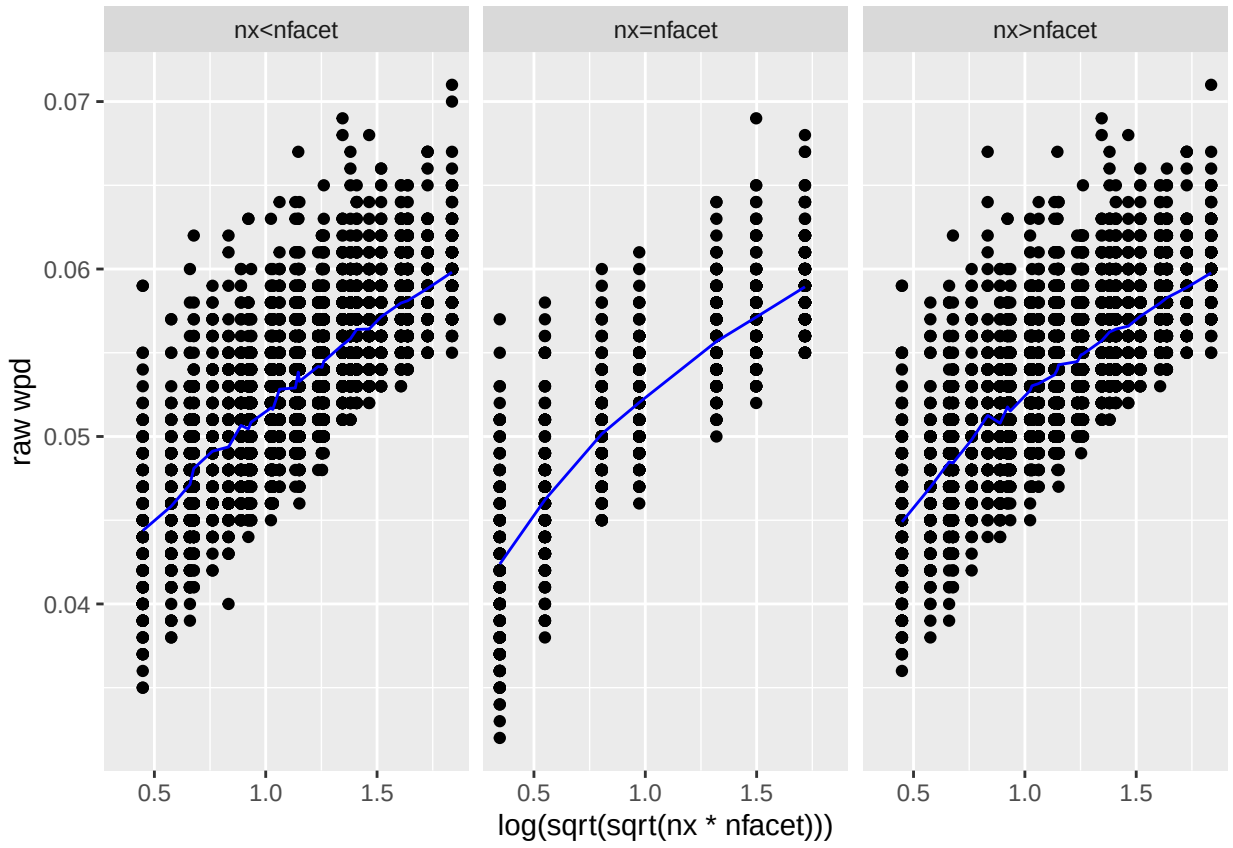
Looks like for all the variations, there is a quadratic relationship between wpd values and nx*nfacet



- Attempt to linearize it with sqrt, sqrt(sqrt) and log (sqrt(sqrt))

The transformation log (sqrt(sqrt())) on nx*nfacet finally makes it approximately linear.

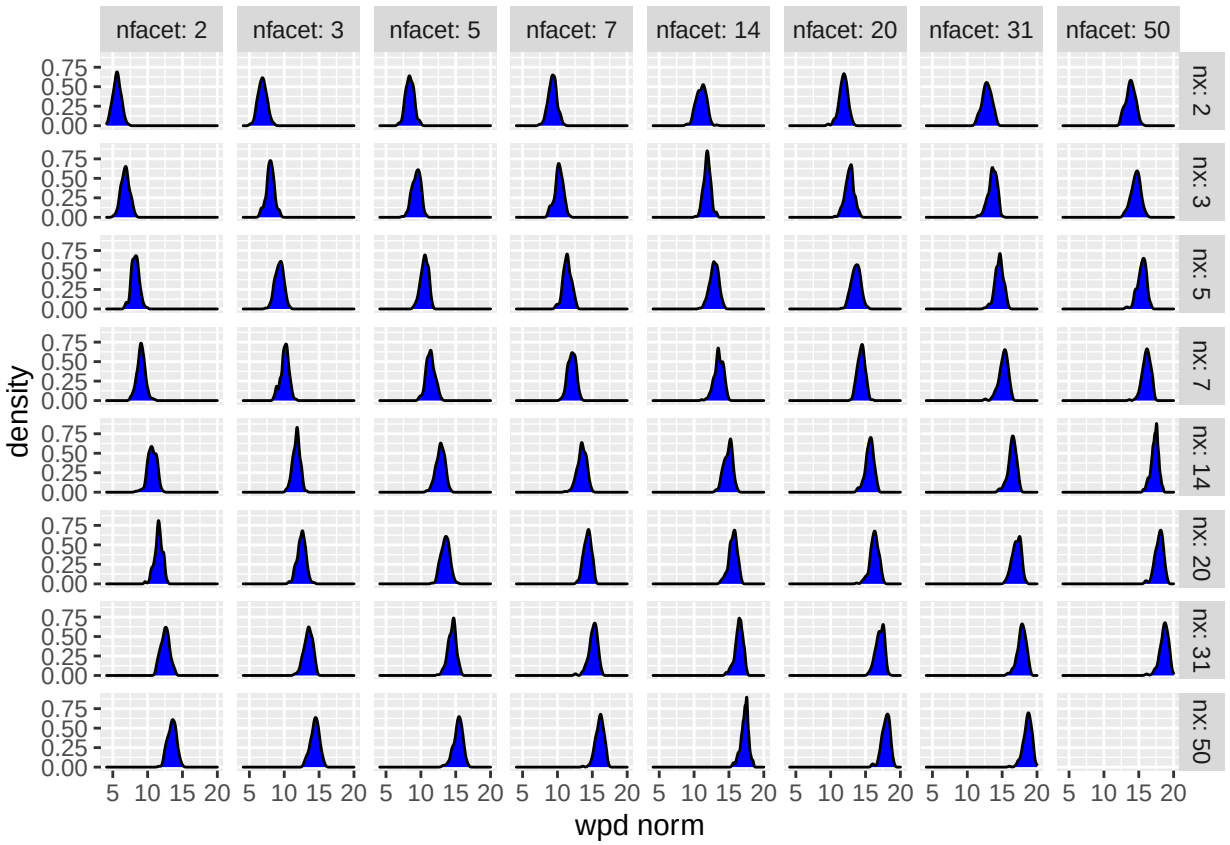

nx<nfacet
nx=nfacet
nx>nfacet
raw wpd
0.07
0.06
0.05
0.04
10
20
30
40
sqrt(nx * nfacet)
nx<nfacet
nx=nfacet
nx>nfacet
raw wpd
0.07
0.06
0.05
0.04
2
3
4
5
6
sqrt(sqrt(nx * nfacet))



- Distribution after linearizing it

The distribution of $wpd_{norm} = log(sqrt(sqrt(nx * nfacet)))/wpd$ is plotted. The shape and spread look similar but location is shifting to the right.

(If we define wpd_{norm} as the inverse of it, the values become too small and the distribution too skewed. Hence the inverse of it is considered.)



What we want is a transformation which will be constant and not linear to obtain similar locations for all panels.