Update on visuals

1. Standardization
2. Severity cutoff

1. Standardization

GMM 1st component standardization code

standardization (gmm)

https://scikit-learn.org/stable/auto_examples/mixture/plot_gmm_covariances.html cov

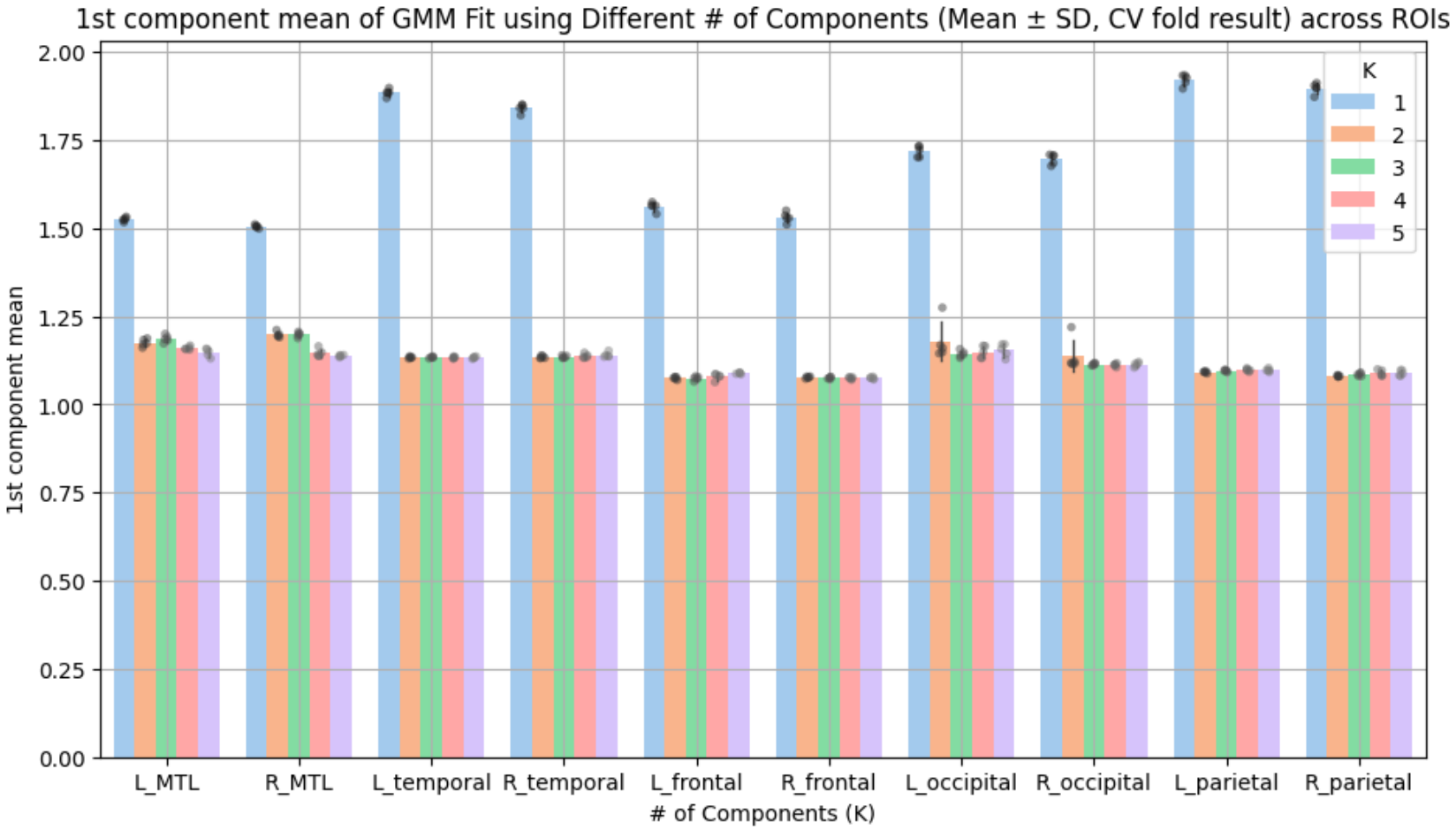
covariance_type{'full', 'tied', 'diag', 'spherical'}, default='full', each component has its own covariance matrix (in 1d just one value)

```
gmm1mean = []
gmm1std = []
dat = pd.read_csv('wide_data.csv').loc[:,lobes].values #read in the data agian
gmm_data = dat.copy()
for roi in lobes:
    roi_idx = lobes.index(roi)
    gmm_model = GaussianMixture(n_components=2, random_state=42)
    # Fit GMM to the column data/a specific roi
    gmm_model.fit(gmm_data[:, roi_idx].reshape(-1, 1))
    # Get the means of the two components
    means = gmm_model.means_.flatten()
    # sort it by smaller mean = first component
    c1_idx = np.argmin(means)
    std = np.sqrt(gmm_model.covariances_[c1_idx])[0][0]
    mean = means[c1_idx]
    gmm1mean.append(mean)
    gmm1std.append(std)
    #standardize
    gmm_data[:, roi_idx] = (gmm_data[:, roi_idx] - mean)/std
```

[144] ✓ 0.1s Python

[GaussianMixture function from sklearn]

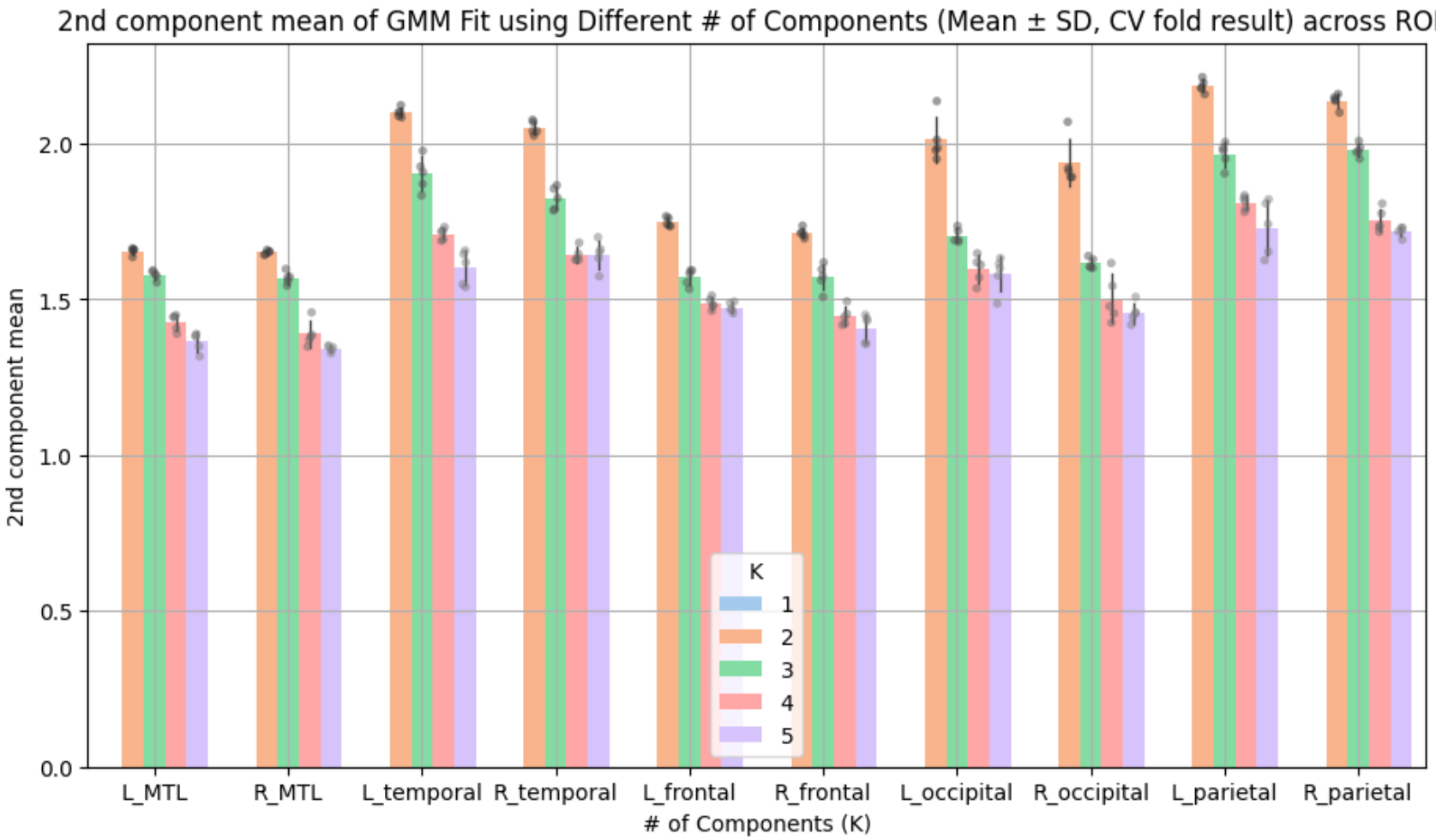
1st component mean from K-GMM fit on mean SUVR



Slight difference in 1st component mean between 2-GMM and 3-GMM fit in occipital ROIs.

Notebook section: ### best # of GMM components

2nd component mean from K-GMM fit on mean SUVR



As expected, affect 2nd component mean more but 2nd component statistics are not involved in standardization.

Notebook section: ### best # of GMM components

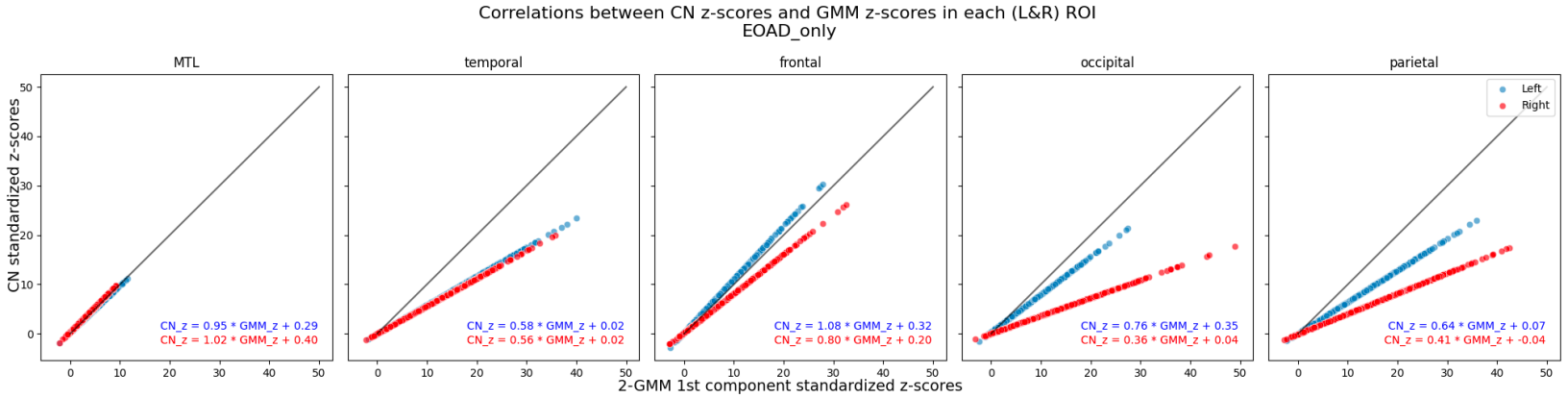
Difference in 1st component mean and sd between CN and GMM z-scores

	ROI	GMM_Mean	CN_Mean	GMM_Std	CN_Std	Mean_Percentage_Diff	Std_Percentage_Diff
0	L_MTL	1.177341	1.147086	0.099673	0.104892	2.637606	-4.975370
1	R_MTL	1.195019	1.149653	0.115304	0.113034	3.945989	2.008232
2	L_temporal	1.133065	1.131258	0.059461	0.101918	0.159729	-41.657859
3	R_temporal	1.133864	1.131780	0.064895	0.115925	0.184168	-44.020064
4	L_frontal	1.075524	1.054900	0.070502	0.065304	1.955086	7.959815
5	R_frontal	1.076495	1.060796	0.062974	0.078968	1.479965	-20.254425
6	L_occipital	1.156955	1.112627	0.096696	0.126598	3.984075	-23.619721
7	R_occipital	1.118323	1.111354	0.057420	0.158739	0.627087	-63.827187
8	L_parietal	1.091655	1.083718	0.075538	0.118149	0.732337	-36.065979
9	R_parietal	1.080651	1.086896	0.063153	0.153422	-0.574567	-58.836766

The percentage difference is calculated as (GMM – CN) / CN

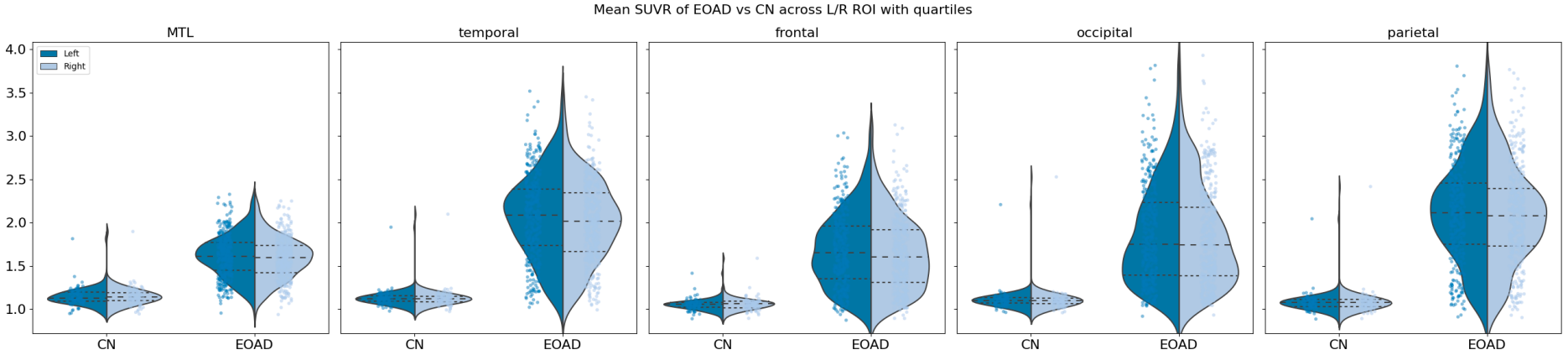
Larger difference in sd values in the temporal, occipital, and parietal ROIs:

GMM standardization gives smaller sd values which agrees with the slopes observed in the correlation graph.



Notebook section: ### percentage diff. w.r.t. cn -> ## correlations between gmm and cn zscores

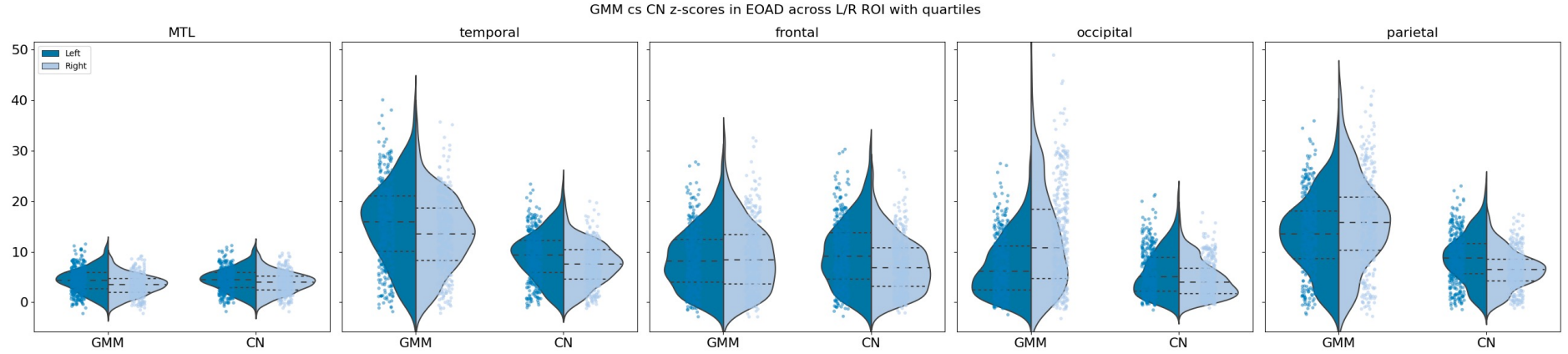
Visualize z-scores and mean SUVR distribution difference



Mean SUVR value distributions seem to be mostly symmetric in L and R across all ROI (see next for combined

In CN, there is a significant tail due to having an outlier in both L and R, but R's outlier value is larger in all ROIs.

In EOAD, L seems to have higher values looking at where the quartile lines are drawn.



This graph compares different z-scores looking at the range of z-score values and symmetry between L & R ROIs:

Not much difference between the two z-scores in MTL and frontal, which is also reflected in slopes ~ 1 and %diff in sd from the previous.

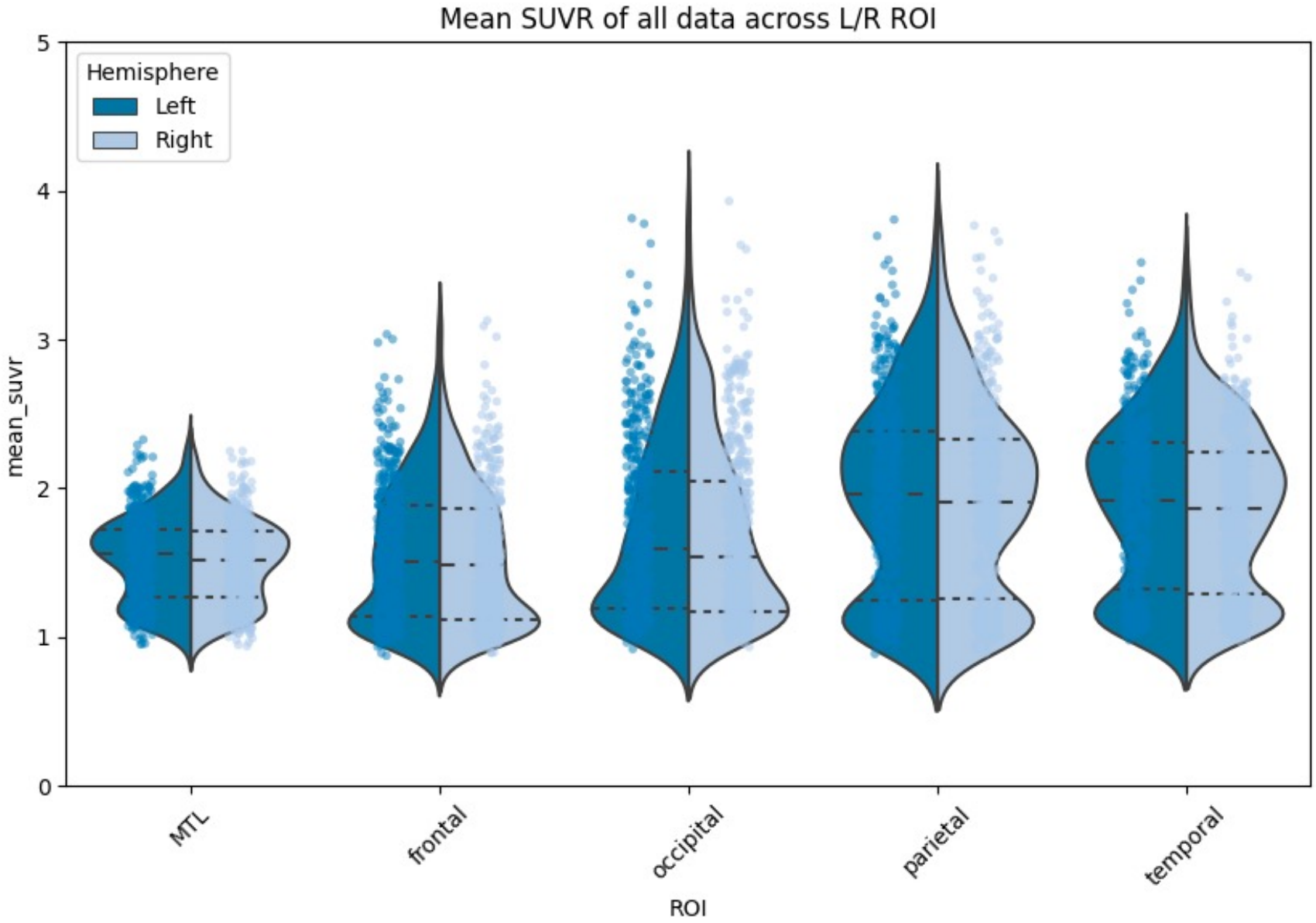
Concerning differences in temporal, occipital, and parietal: 1) GMM z-scores have much wider ranges due to the diff. in sd;

2) while CN z-scores seem to preserve symmetry (and slightly higher values in L), this is not the case for occipital and parietal (slopes deviate from each other). Looking at the strip plots, R ends up having higher values, especially in occipital.

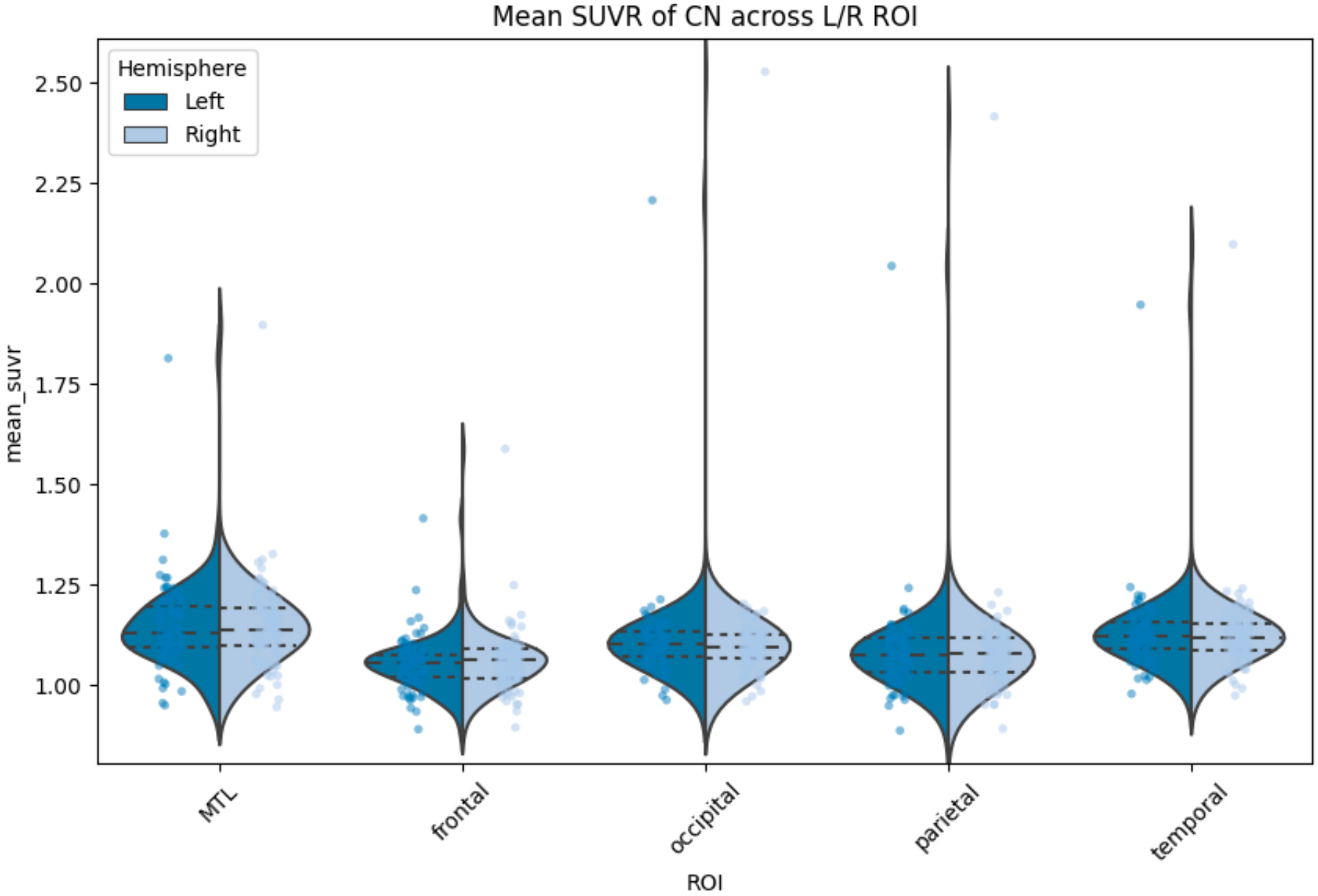
If the codes are right I think the only explanation is different sd being calculated which might be affected by CN outliers.

Notebook section: ## Visualize L/R value difference and ### violin compare z-scoring method

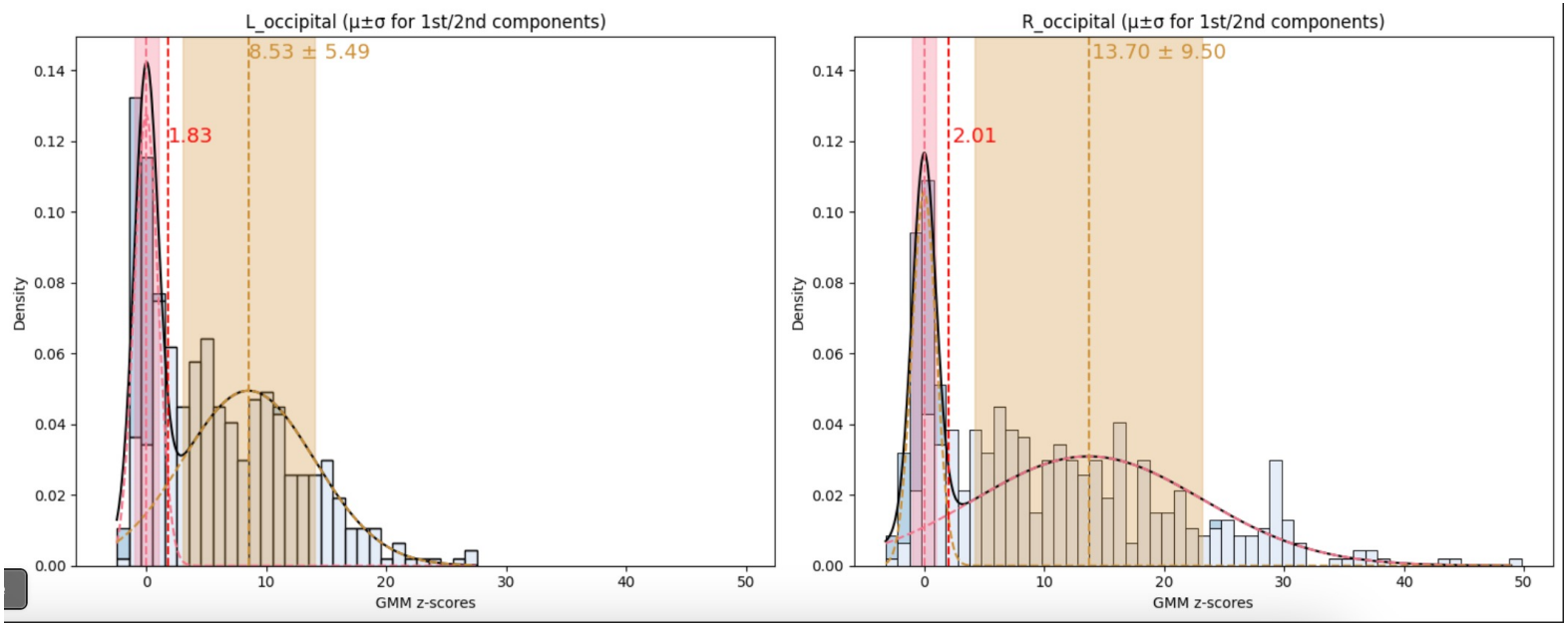
All data

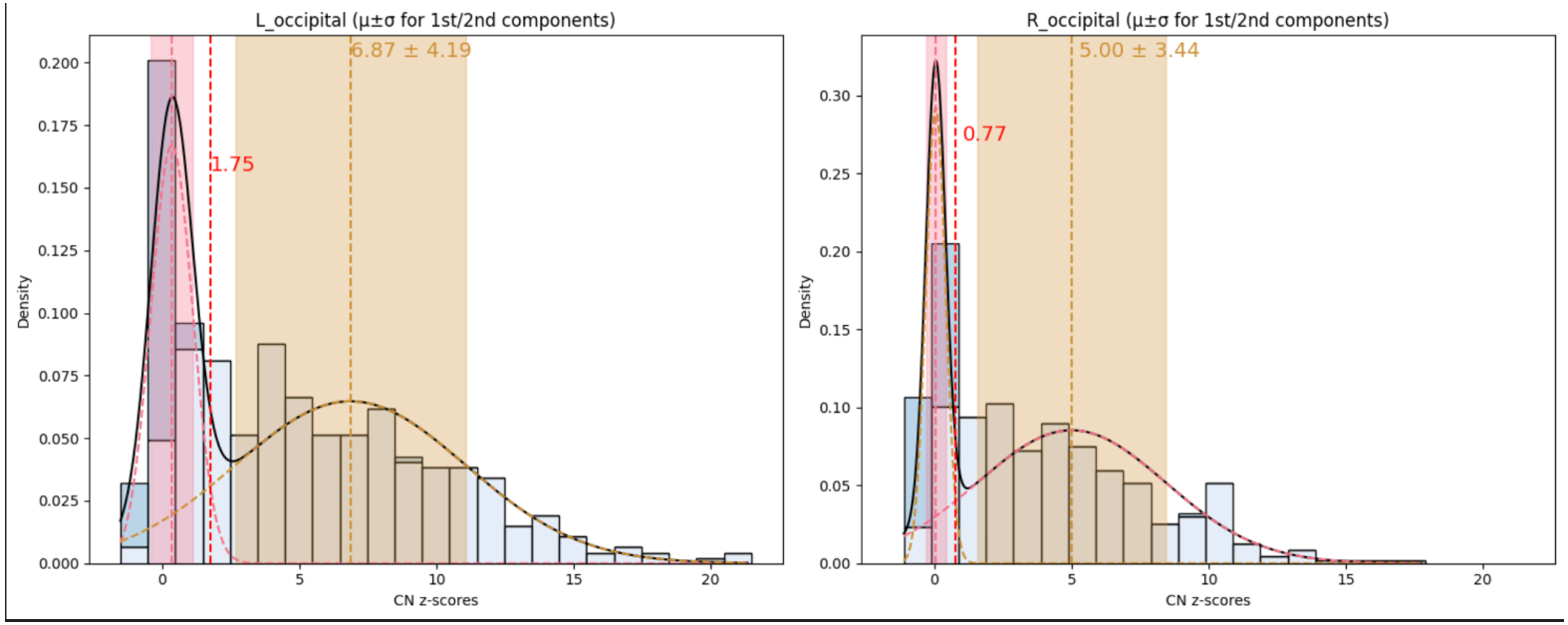

Mean SUVR of all data across L/R ROI
Hemisphere
Left
Right
mean_suvr
0
1
2
3
4
5
MTL
frontal
occipital
parietal
temporal
ROI

CN only


Mean SUVR of CN across L/R ROI
Hemisphere
Left
Right
mean_suvr
2.50
2.25
2.00
1.75
1.50
1.25
1.00
MTL
frontal
occipital
parietal
temporal
ROI

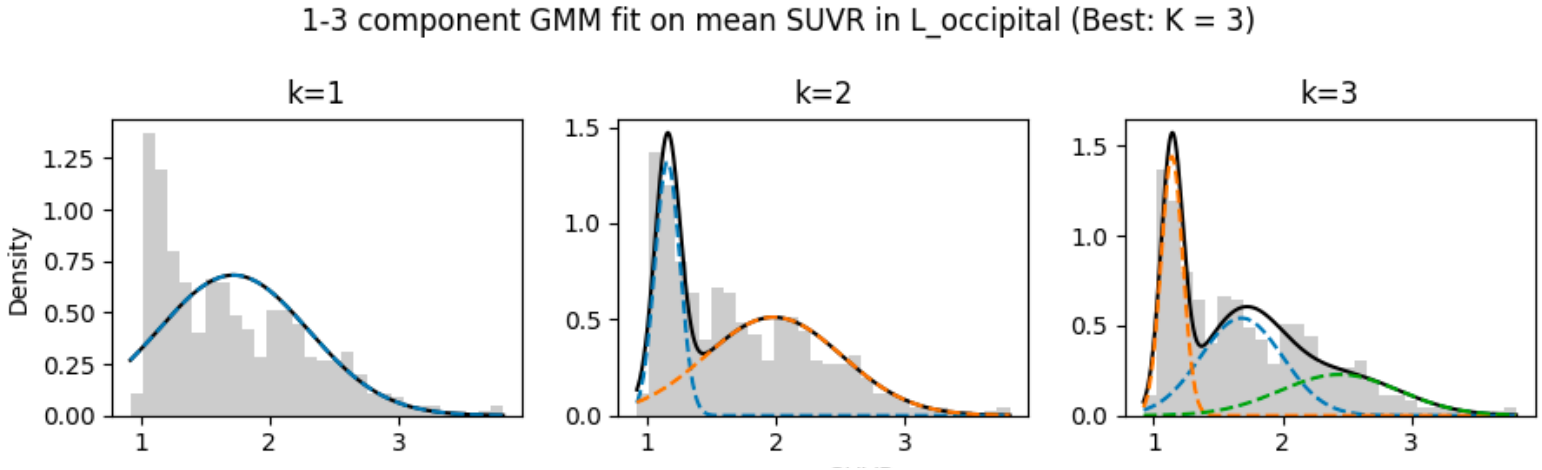
Occipital

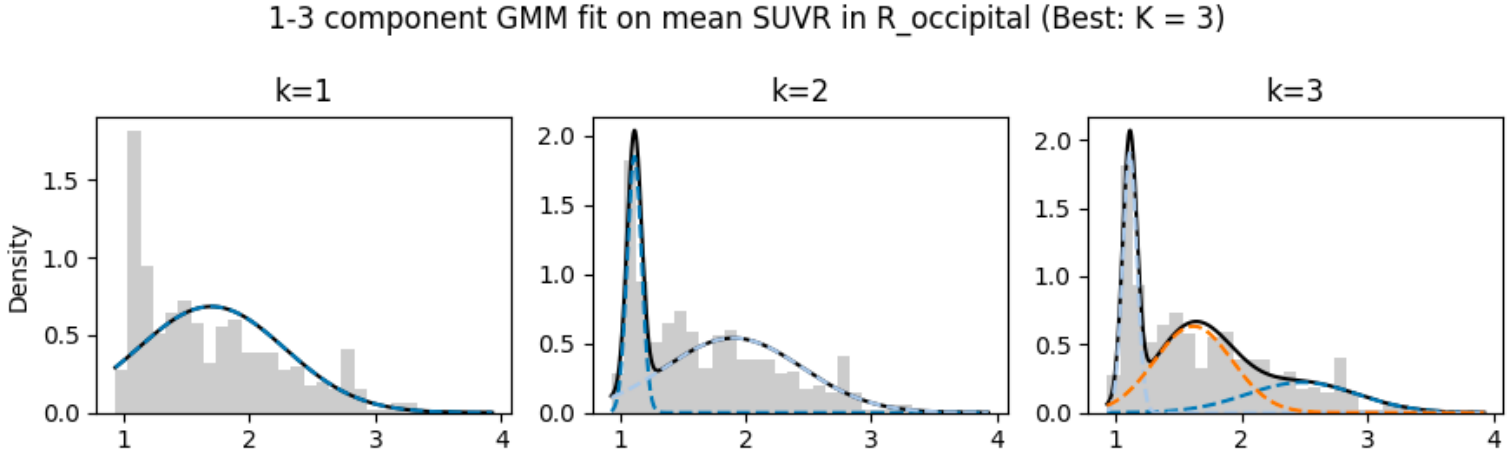

L_occipital (μ±σ for 1st/2nd components)
8.53 ± 5.49
1.83
Density
GMM z-scores
R_occipital (μ±σ for 1st/2nd components)
13.70 ± 9.50
2.01
Density
GMM z-scores


L_occipital (μ±σ for 1st/2nd components)
6.87 ± 4.19
1.75
Density
CN z-scores
R_occipital (μ±σ for 1st/2nd components)
5.00 ± 3.44
0.77
Density
CN z-scores

if 3-GMM were to be fitted on occipital mean SUVR the result is similar

Recall from CV AIC occipital ROIs are slightly better fitted with 3-GMM
(looking at the histograms instead of the overall density function)
Notebook section: ### func: k-GMM fit in a specific ROI





	ROI	GMM_Mean	CN_Mean	GMM_Std	CN_Std	Mean_Percentage_Diff	Std_Percentage_Diff
6	L_occipital	1.142589	1.112627	0.085631	0.126598	2.692823	-32.359833
7	R_occipital	1.114066	1.111354	0.056022	0.158739	0.243999	-64.707915

2. Cutoff

Levels derived using intersection, 2nd component mean ± sd on cn/gmm z-scores

```
print('levels on gmm z-scores')
calculate_gmm_cutoffs(gmm_data)
```

✓ 0.1s

levels on gmm z-scores

	roi	intersection	c2mean-sd	c2mean	c2mean+sd
0	L_MTL	1.43	2.65	4.80	6.96
1	R_MTL	1.32	2.17	3.93	5.69
2	L_temporal	2.24	8.68	16.23	23.77
3	R_temporal	2.13	7.40	14.08	20.76
4	L_frontal	1.87	3.95	9.55	15.14
5	R_frontal	1.87	3.80	10.09	16.38
6	L_occipital	1.83	3.04	8.53	14.02
7	R_occipital	2.01	4.20	13.70	23.20
8	L_parietal	2.17	7.63	14.46	21.29
9	R_parietal	2.22	8.38	16.73	25.08

```
print('levels on cn z-scores')
calculate_gmm_cutoffs(zdata)
```

✓ 0.2s

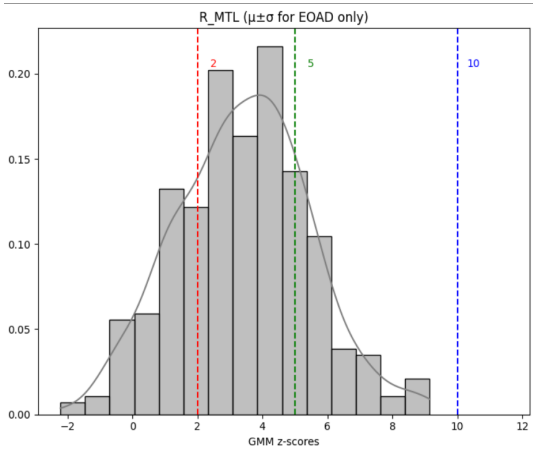
levels on cn z-scores

	roi	intersection	c2mean-sd	c2mean	c2mean+sd
0	L_MTL	1.65	2.81	4.85	6.90
1	R_MTL	1.75	2.61	4.41	6.20
2	L_temporal	1.33	5.08	9.48	13.89
3	R_temporal	1.21	4.16	7.90	11.64
4	L_frontal	2.33	4.59	10.62	16.66
5	R_frontal	1.69	3.23	8.25	13.26
6	L_occipital	1.75	2.67	6.87	11.06
7	R_occipital	0.77	1.56	5.00	8.43
8	L_parietal	1.46	4.94	9.31	13.68
9	R_parietal	0.87	3.41	6.85	10.28

Notebook section: ## varied. but functions are defined above the section

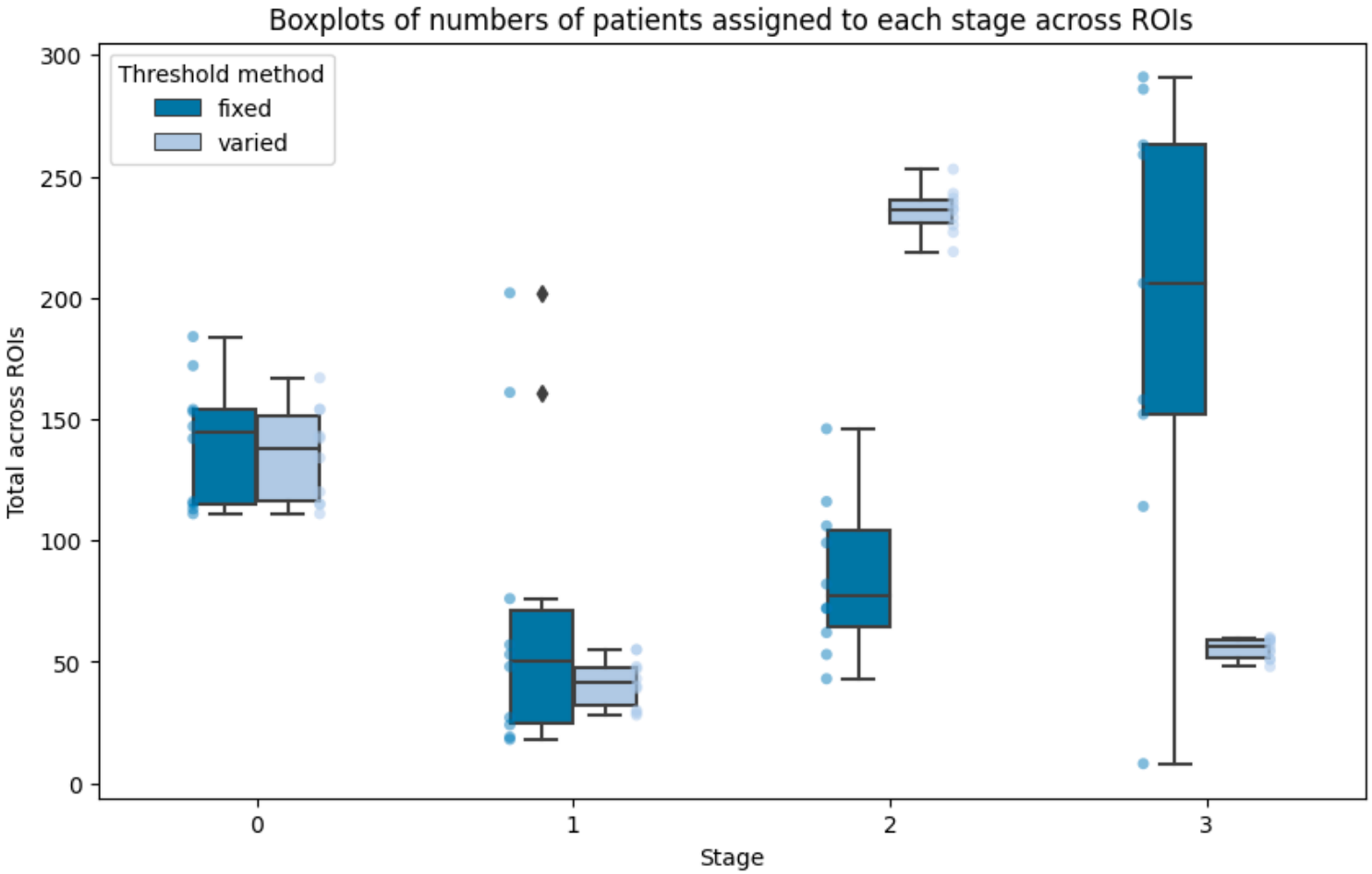
Some visualizations

- The following slides intend to visualize the difference in the number of subjects assigned to each stage by different ROI and cutoff selection methods
- cutoff selection methods being: fixed (= 2, 5, 10) v.s. varied (= intersection, 2nd component mean $\pm$ sd in each ROI), on GMM 1st comp. standardized z-scores
- can plot with other z-scores or levels but I think it's worth figuring out what's happening with the GMM z-scores first
- the stages are named 0, 1, 2, and 3 separated by the three levels
- note that in R_MTL no stage 3 (beyond the 3rd level) is assigned for the fixed method because 10 is beyond the maximum z-scores in that ROI
  (same in CN z-scores)

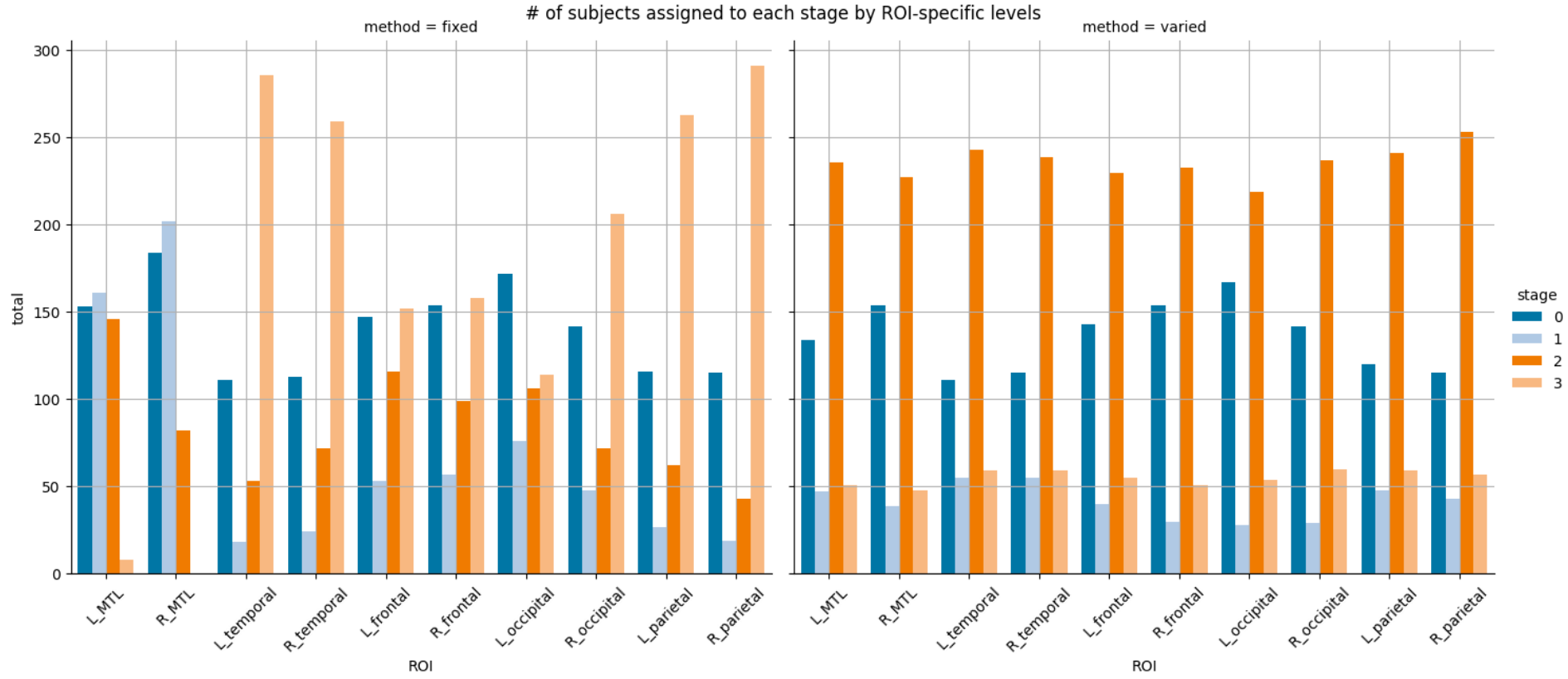


Notebook section: # Difference due to cutoff choice

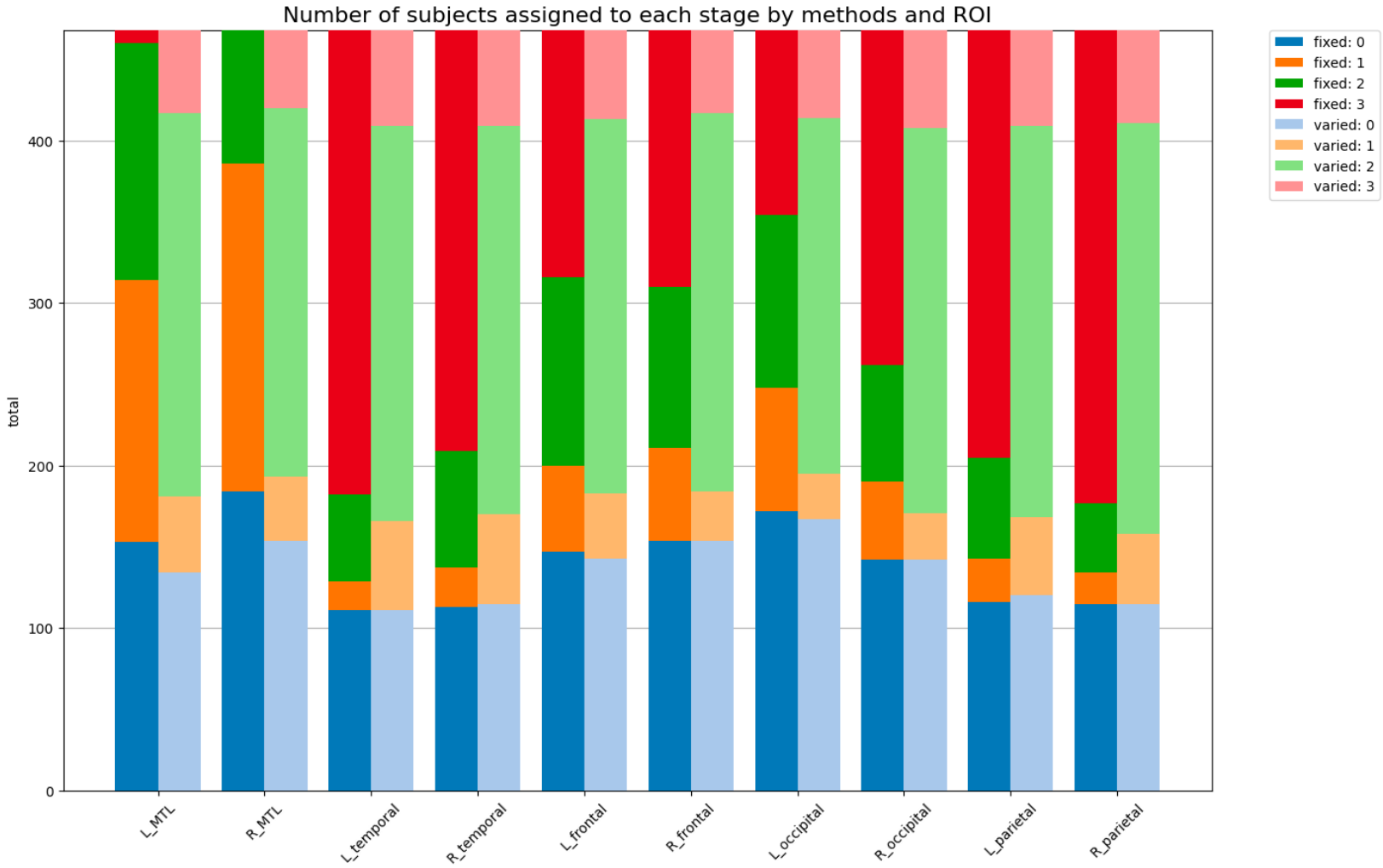
Boxplots


Boxplots of numbers of patients assigned to each stage across ROIs
Threshold method
fixed
varied
Total across ROIs
300
250
200
150
100
50
0
0
1
2
3
Stage

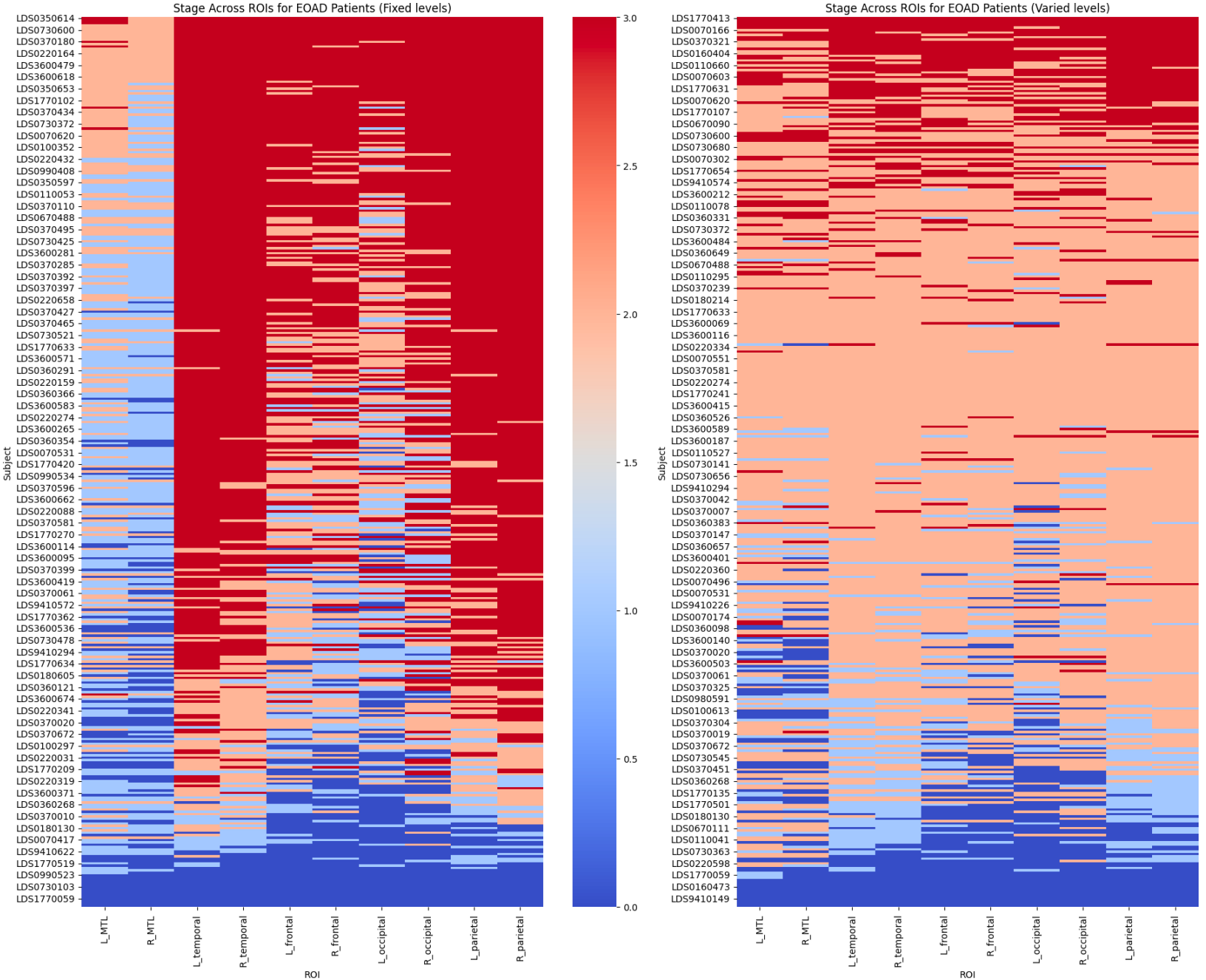
Clustered bar



Clustered stacked bar



heatmap


Stage Across ROIs for EOAD Patients (Fixed levels)
Subject
ROI
L_MTL
R_MTL
L_temporal
R_temporal
L_frontal
R_frontal
L_occipital
R_occipital
L_parietal
R_parietal
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Stage Across ROIs for EOAD Patients (Varied levels)
Subject
ROI

Thoughts

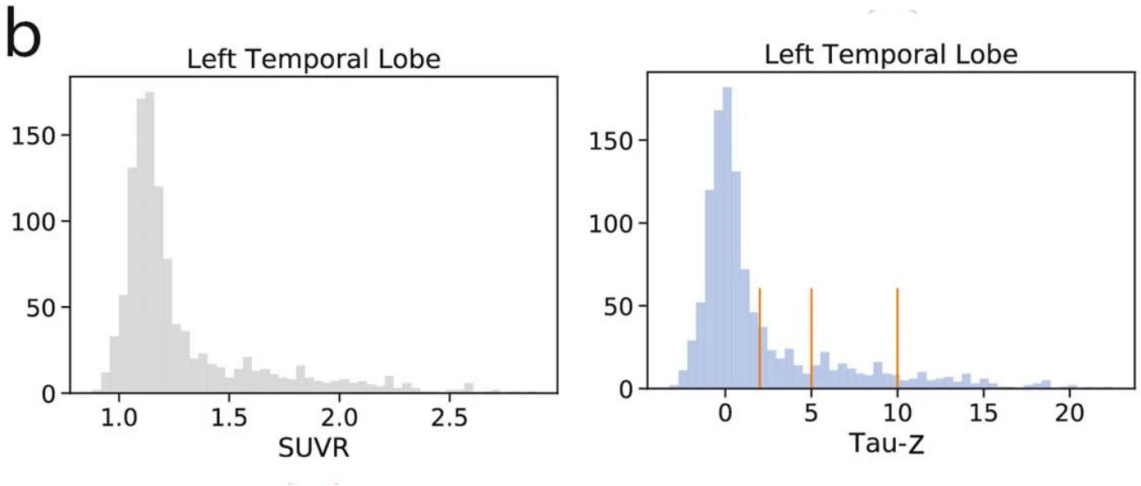
I think it's apparent from these graphs (also reflected in the individual ROI's component density functions + cutoff vertical lines graph):

- Varied (from mean and sd of the 2nd component) assigns most to stage 2, which is between mean – sd and mean + sd. Expected as that range does incorporate a large percentage of the EOAD data; in contrast to stage 0 which is meant for the noise (normal) component, and stage 1 which is a much narrower range due to the intersection of the two components being close to the mean – sd of the 2nd component.
- Fixed (2, 5, 10) assigns most to stage 3, which is past the final severity level. This is expected because 10 is smaller than the 2nd component mean in most ROIs when a 2-GMM is fitted on the GMM z-scores.

- Then, this circles back to the question of how exactly is 2, 5, and 10 chosen:
  - In Vogel et al. I don't think there is any more discussion aside from what's talked about.
  - In Young et al. I think they start with four levels, where the first three are also arbitrarily chosen as 1, 2, 3. But if beyond the 3rd level less than 10 subjects were observed then the number of levels is reduced. The maximum z-score is set to be 2, 3, or 5 depending on how many levels there already are. (This is discussed in the SuStaIn modeling of GENFI/ADNI and supp. Table 6.) I have no idea of the data distribution, and the z-scores are derived by standardizing against the control.

Compare

a figure showing the mean SUVR and z-scores for the left temporal ROI from Vogel et al's : (to compare I've adjusted stat=count and binwidth in example z-scoring).

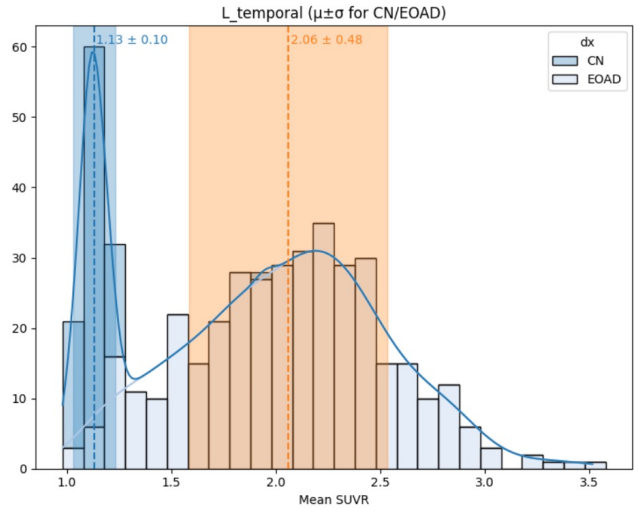


(For L_temporal: = parahippocampal, inferiortemporal, fusiform, middletemporal,superiortemporal, transversetemporal)

Mean SUVR has a wider range in ours (and more bi-modal/separated CN vs EOAD?) which after transformation becomes an even wider range of z-scores.

← The final level of 10 seems to be the midpoint of the 'Tau-Z' here.

But in our z-scores 10 is roughly the 2nd level.



Fixed levels