Testing out scaling

cool people

Get a baseline (centered X, centered/scaled Z)

```
# Load in simulation studies

# Not doing the p100 n300 case since a single simulation
# requires over 16 hours on the cluster and many simulations
# need to be done

p = c(5, 15, 25, 50)
n = c(90, 90, 150, 150)
objects_strings = c(
  "simulation_study//p5_n90//res_p5_n90_covdepGE_20220908_215120.Rda",
  "simulation_study//p15_n90//res_p15_n90_covdepGE_20220908_215229.Rda",
  "simulation_study//p25_n150//res_p25_n150_covdepGE_20220825_121750.Rda",
  "simulation_study//p50_n150//res_p50_n150_covdepGE_20220825_090326.Rda"
  # "simulation_study//p100_n300//res_p100_n300_covdepGE_20220824_084919.Rda"
)

results_original = list()
for(sim in 1:length(objects_strings)) {
  load(objects_strings[sim])
  results_original[[sim]] = results
  rm(results)
}
sim_names_original = paste0("p", p, "_n", n)
results_original = set_names(results_original, sim_names_original)
```

Calculate mean FP/n and FN/n

```r
false_positives_baseline = results_original %>%
  map(function(x)
    map_dbl(x, pluck, "FP_n")
  )
false_negatives_baseline = results_original %>%
  map(function(x)
    map_dbl(x, pluck, "FN_n")
  )

false_positives_baseline %>%
  map_dfr(summary) %>%
  cbind(p, n, .) %>%
  tibble() %>%
  xtable(caption = "False positives per sample - Normalized Z, Centered X")
```

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Mon Nov 14 12:57:01 2022

	p	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5.00	90.00	0.00	0.00	0.03	0.34	0.47	2.62
2	15.00	90.00	0.00	0.00	0.62	0.82	1.25	4.42
3	25.00	150.00	0.00	0.64	1.08	1.29	1.75	5.17
4	50.00	150.00	0.39	2.68	4.37	4.36	5.80	11.31

Table 1: False positives per sample - Normalized Z, Centered X

```r
false_negatives_baseline %>%
  map_dfr(summary) %>%
  cbind(p, n, .) %>%
  tibble() %>%
  xtable("False negatives per sample - Normalized Z, Centered X")
```

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Mon Nov 14 12:57:01 2022

	p	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5.00	90.00	0.00	0.78	1.03	0.99	1.29	2.16
2	15.00	90.00	0.00	0.89	1.32	1.41	1.99	2.98
3	25.00	150.00	0.00	0.77	0.92	0.91	1.10	2.00
4	50.00	150.00	0.08	0.92	1.13	1.19	1.40	3.28

Table 2: False negatives per sample - Normalized Z, Centered X

Goal: reduce mean FP/n (and either reduce or keep constant FN/n)

Data generation

```
set.seed(12345)
n_trials = 100

simulation_list = map2(n, p, function(n,p){
  nj = n %/% 3
  replicate(n_trials, generateData(p, nj, nj, nj), F)
})
```

```
max_min_scale = function(X) { # Scale each column by max-min
  # Columnwise; subtract min and divide by resulting max
  if(!is.matrix(X)) { # for vectors
    X = as.matrix(X)
  }
  p = ncol(X)
  n = nrow(X)
  # faster with max.col(X) and max.col(-X)
  mins = as.vector(apply(X, 2, min))
  scaled_X = t(t(X)-mins)
  maxs = as.vector(apply(scaled_X, 2, max))
  scaled_X = t(t(scaled_X)/maxs)
  return(list(scaled_X = scaled_X, add_invs = mins, mult_invs = maxs))
}

max_min_unscale = function(output) {
  t(t(output$scaled_X)*output$mult_invs + output$add_invs)
}
```

Raw performance (no center/scaling)

Test max-min scaling

We'll try 3 situations; scaling both X and Z, scaling only Z, and scaling only X

```
min_max_simulation_list = map(simulation_list, function(setup){
  mm_X_data_list = map(setup, function(sim) {
    output = max_min_scale(sim$X)
```

```
    sim$X = output$scaled_X
    sim$scaleoutX = output
    sim
  })
  mm_Z_data_list = map(setup, function(sim) {
    output = max_min_scale(sim$Z)
    sim$Z = output$scaled_X
    sim$scaleoutZ = output
    sim
  })
  mm_XZ_data_list = map(setup, function(sim) {
    output = max_min_scale(sim$X)
    sim$X = output$scaled_X
    sim$scaleout = output
    output2 = max_min_scale(sim$Z)
    sim$Z = output2$scaled_X
    sim$scaleoutZ = output2
    sim
  })
  list(mm_X_data_list, mm_Z_data_list, mm_XZ_data_list)
})
```

```
num_workers <- parallel::detectCores() - 8
doParallel::registerDoParallel(cores = num_workers)

min_max_X_simulation_results = min_max_simulation_list %>% map(~pluck(.x, 1)) %>%
  map(function(setup){
    simulation_func(n_trials, setup, num_workers, normalize = FALSE)
  })

save(min_max_X_simulation_results, file = "minmax_X_sim.Rda")

min_max_Z_simulation_results = min_max_simulation_list %>% map(~pluck(.x, 2)) %>%
  map(function(setup){
    simulation_func(n_trials, setup, num_workers, normalize = FALSE)
  })

save(min_max_Z_simulation_results, file = "minmax_Z_sim.Rda")

min_max_XZ_simulation_results = min_max_simulation_list %>% map(~pluck(.x, 3)) %>%
  map(function(setup){
```

```
    simulation_func(n_trials, setup, num_workers, normalize = FALSE)
  })

save(min_max_XZ_simulation_results, file = "minmax_XZ_sim.Rda")
```

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Mon Nov 14 12:57:02 2022

	p	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5.00	90.00	0.00	0.37	0.64	0.81	1.19	2.53
2	15.00	90.00	0.78	3.03	4.13	4.08	4.78	8.49
3	25.00	150.00	7.57	10.16	11.87	11.89	13.53	19.25
4	50.00	150.00	11.09	17.41	20.21	19.86	22.22	28.53

Table 3: False positives per sample - Max/Min Scaled Z

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Mon Nov 14 12:57:02 2022

	p	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5.00	90.00	0.11	0.80	1.16	1.14	1.42	2.42
2	15.00	90.00	0.71	1.73	2.36	2.35	2.83	4.07
3	25.00	150.00	0.32	1.20	1.66	1.71	2.13	4.19
4	50.00	150.00	1.07	1.92	2.27	2.42	2.84	4.08

Table 4: False negatives per sample - Max/Min Scaled Z

Test Max-min + Normalization

First do a max-min transform to scale, then do the z-transform (or mean 0 center transform in the case of X)

```
num_workers <- parallel::detectCores() - 8
doParallel::registerDoParallel(cores = num_workers)

min_max_norm_X_simulation_results = min_max_simulation_list %>% map(~pluck(.x, 1)) %>%
  map(function(setup){
    simulation_func(n_trials, setup, num_workers, normalize = TRUE)
  })

save(min_max_norm_X_simulation_results, file = "minmax_norm_X_sim.Rda")

min_max_norm_Z_simulation_results = min_max_simulation_list %>% map(~pluck(.x, 2)) %>%
  map(function(setup){
```

```
    simulation_func(n_trials, setup, num_workers, normalize = TRUE)
  })

save(min_max_norm_Z_simulation_results, file = "minmax_norm_Z_sim.Rda")

min_max_norm_XZ_simulation_results = min_max_simulation_list %>% map(~pluck(.x, 3)) %>%
  map(function(setup){
    simulation_func(n_trials, setup, num_workers, normalize = TRUE)
  })

save(min_max_norm_XZ_simulation_results, file = "minmax_norm_XZ_sim.Rda")
```

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Mon Nov 14 12:57:05 2022

	p	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5.00	90.00	0.00	0.00	0.00	0.26	0.34	2.51
2	15.00	90.00	0.00	0.00	0.33	0.61	0.96	3.89
3	25.00	150.00	0.00	0.64	1.09	1.34	1.97	3.73
4	50.00	150.00	0.55	2.81	4.04	4.07	5.11	10.25

Table 5: False positives per sample - Max/Min Scaled X and Normalization

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Mon Nov 14 12:57:05 2022

	p	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5.00	90.00	0.00	0.78	0.96	1.00	1.33	1.98
2	15.00	90.00	0.00	1.11	1.42	1.48	2.00	2.87
3	25.00	150.00	0.00	0.72	0.97	0.92	1.15	1.96
4	50.00	150.00	0.16	0.87	1.06	1.12	1.32	2.67

Table 6: False negatives per sample - Max/Min Scaled X and Normalization

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Mon Nov 14 12:57:05 2022

	p	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5.00	90.00	0.00	0.00	0.00	0.24	0.27	2.60
2	15.00	90.00	0.00	0.00	0.36	0.63	0.94	5.04
3	25.00	150.00	0.00	0.60	1.10	1.34	1.99	3.80
4	50.00	150.00	0.32	2.74	3.88	4.08	5.23	10.27

Table 7: False positives per sample - Max/Min Scaled Z and Normalization

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Mon Nov 14 12:57:05 2022

	p	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5.00	90.00	0.00	0.80	0.94	1.03	1.41	1.93
2	15.00	90.00	0.04	1.11	1.49	1.53	2.02	3.09
3	25.00	150.00	0.00	0.72	0.97	0.95	1.24	1.97
4	50.00	150.00	0.19	0.88	1.08	1.15	1.36	2.67

Table 8: False negatives per sample - Max/Min Scaled Z and Normalization

	p	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5.00	90.00	0.00	0.00	0.00	0.24	0.34	1.89
2	15.00	90.00	0.00	0.00	0.33	0.61	0.98	3.89
3	25.00	150.00	0.00	0.65	1.09	1.35	1.97	3.73
4	50.00	150.00	0.55	2.86	3.92	4.07	5.13	10.52

Table 9: False positives per sample - Max/Min Scaled X and Z and Normalization

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Mon Nov 14 12:57:05 2022

% latex table generated in R 4.2.1 by xtable 1.8-4 package % Mon Nov 14 12:57:05 2022

	p	n	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
1	5.00	90.00	0.00	0.78	0.94	1.02	1.33	1.93
2	15.00	90.00	0.02	1.08	1.41	1.47	2.00	2.87
3	25.00	150.00	0.00	0.72	0.95	0.92	1.18	1.96
4	50.00	150.00	0.17	0.85	1.05	1.11	1.31	2.67

Table 10: False negatives per sample - Max/Min Scaled X and Z and Normalization