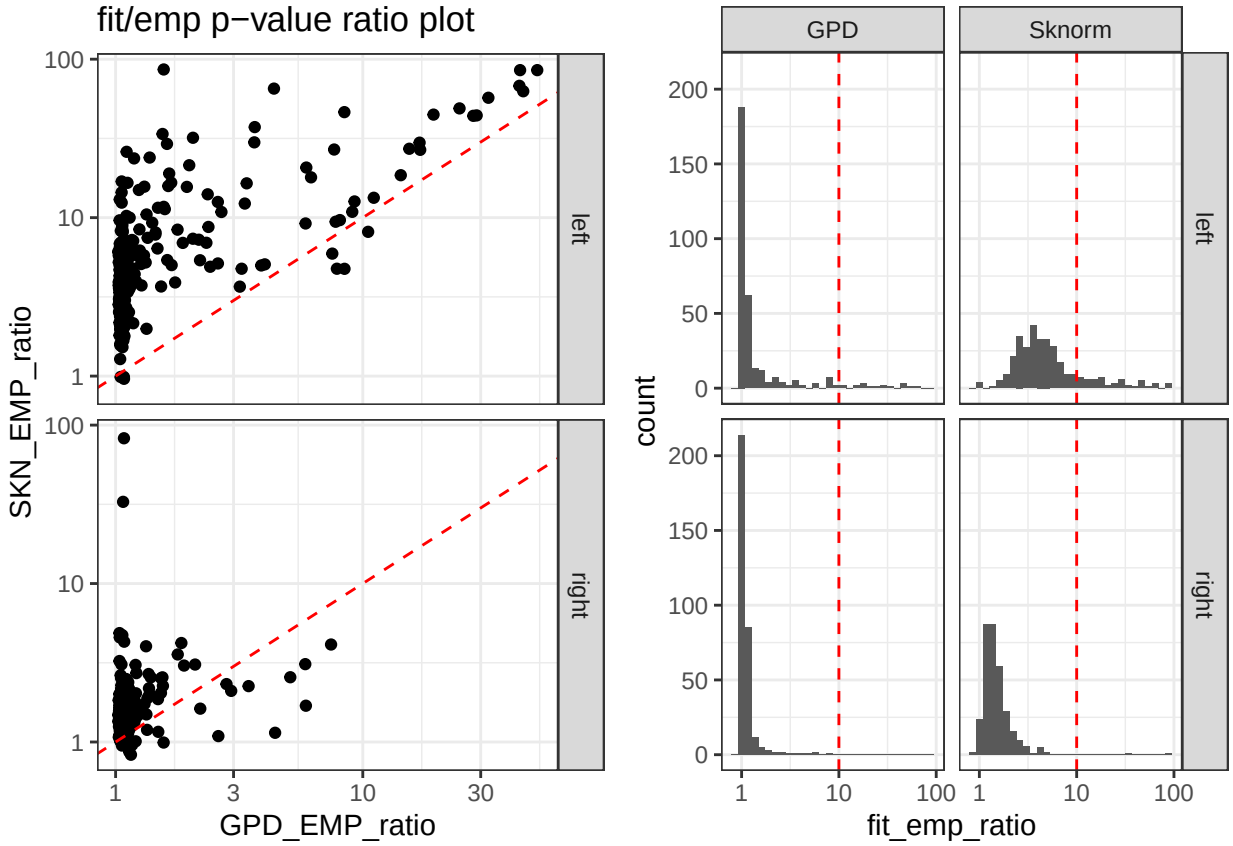
GPD vs Skew Normal

```
library(eva)
library(tidyverse)
library(katlabutils)
library(cowplot)
library(ggplot2)
library(kableExtra)

# should work under low_moi folder
q <- 0.96
samples <- 5e5

# get the file path for GPD and skew normal parameters
subfld <- "figures/power_exploration"
subDir_gpd <- sprintf("%s/tail_prob_%d_resamples_%.2f_percentile", subfld,  samples, q)
subDir_sknorm <- sprintf("%s/sknorm_tail_prob_%d_resamples_%.2f_percentile", subfld, samples, q)
param_gpd <- read_csv(sprintf("%s/param_twosides.csv", subDir_gpd))
param_sknorm <- read_csv(sprintf("%s/param_twosides.csv", subDir_sknorm))
gpd_param <- t(param_gpd[,-1])
sknorm_param <- t(param_sknorm[,-1])


# create the boxplot tibble
param <- tibble(method = c(rep("GPD", 660), rep("Sknorm", 660)),
                tail = rep(c(rep("left", 330), rep("right", 330)), 2),
                GoF_statistic = c(gpd_param[, 4], sknorm_param[, 4]),
                GoF_pvalue = c(gpd_param[, 5], sknorm_param[, 5]),
                fit_emp_ratio = c(gpd_param[, 6], sknorm_param[, 6]),
                emp_fit_ratio = c(gpd_param[, 7], sknorm_param[, 7]))

# create the dotplot tibble
param_comparison <- cbind(param[1:660, 2:6], param[661:1320, 3:6])
colnames(param_comparison)[2:9] <- c("GPD_statistic", "GPD_pvalue",
                                     "GPD_EMP_ratio", "EMP_GPD_ratio",
                                     "SKN_statistic", "SKN_pvalue",
                                     "SKN_EMP_ratio", "EMP_SKN_ratio")
```

1. fit_over_emp ratio plot

```
dot_plot <- param_comparison |>
  filter(GPD_EMP_ratio < 100 & SKN_EMP_ratio < 100) |>
  ggplot(aes_string(x = "GPD_EMP_ratio", y = "SKN_EMP_ratio")) +
  facet_grid(tail ~.) +
  geom_point() +
  geom_abline(linetype = "dashed", color = "red") +
  scale_x_log10() +
  scale_y_log10() +
```

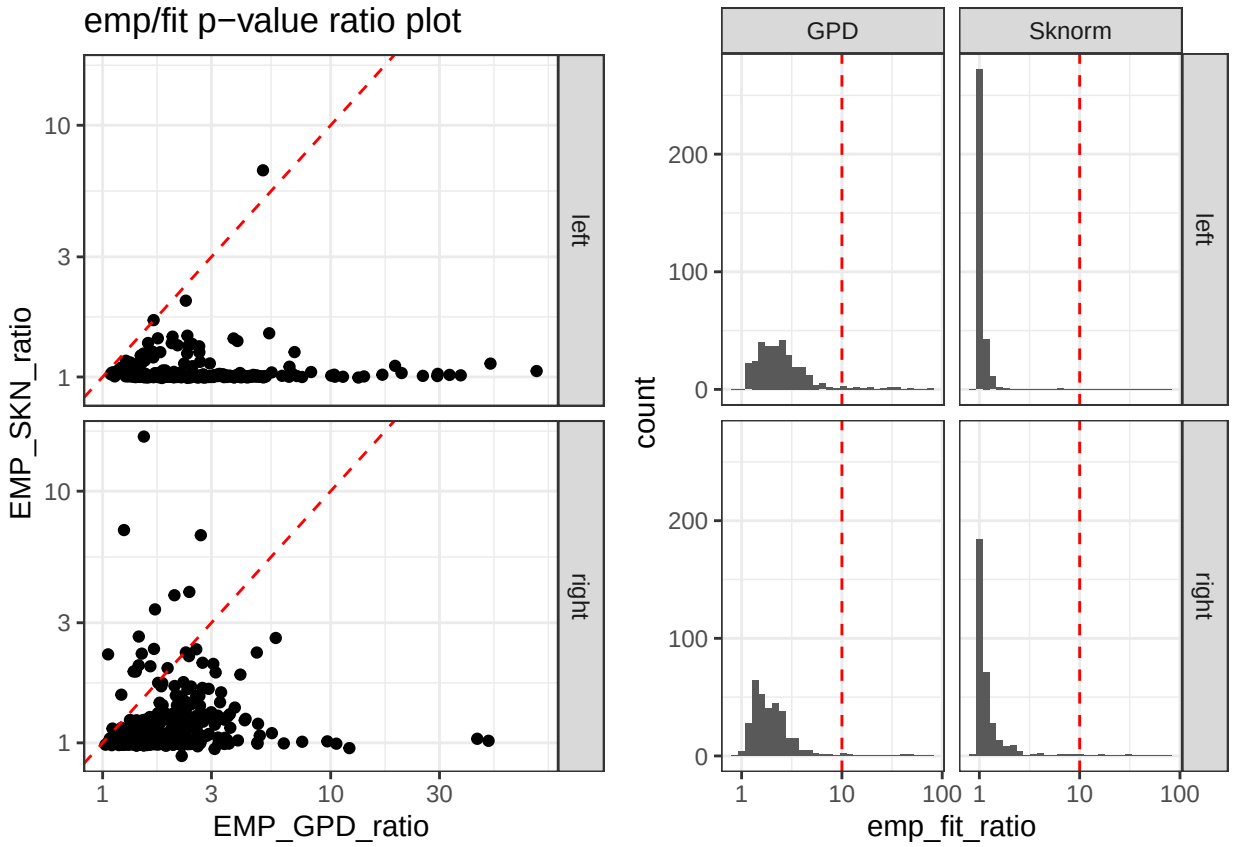
```
  labs(title = "fit/emp p-value ratio plot")

hist_plot <- param |>
  filter(fit_emp_ratio < 100) |>
  ggplot(aes_string(x = "fit_emp_ratio")) +
  facet_grid(tail ~ method) +
  scale_x_log10() +
  geom_histogram() +
  geom_vline(xintercept = 10, linetype = "dashed", color = "red")
plot_grid(dot_plot,
          hist_plot,
          ncol = 2,
          align = "v")
```



2. emp_over_fit ratio plot

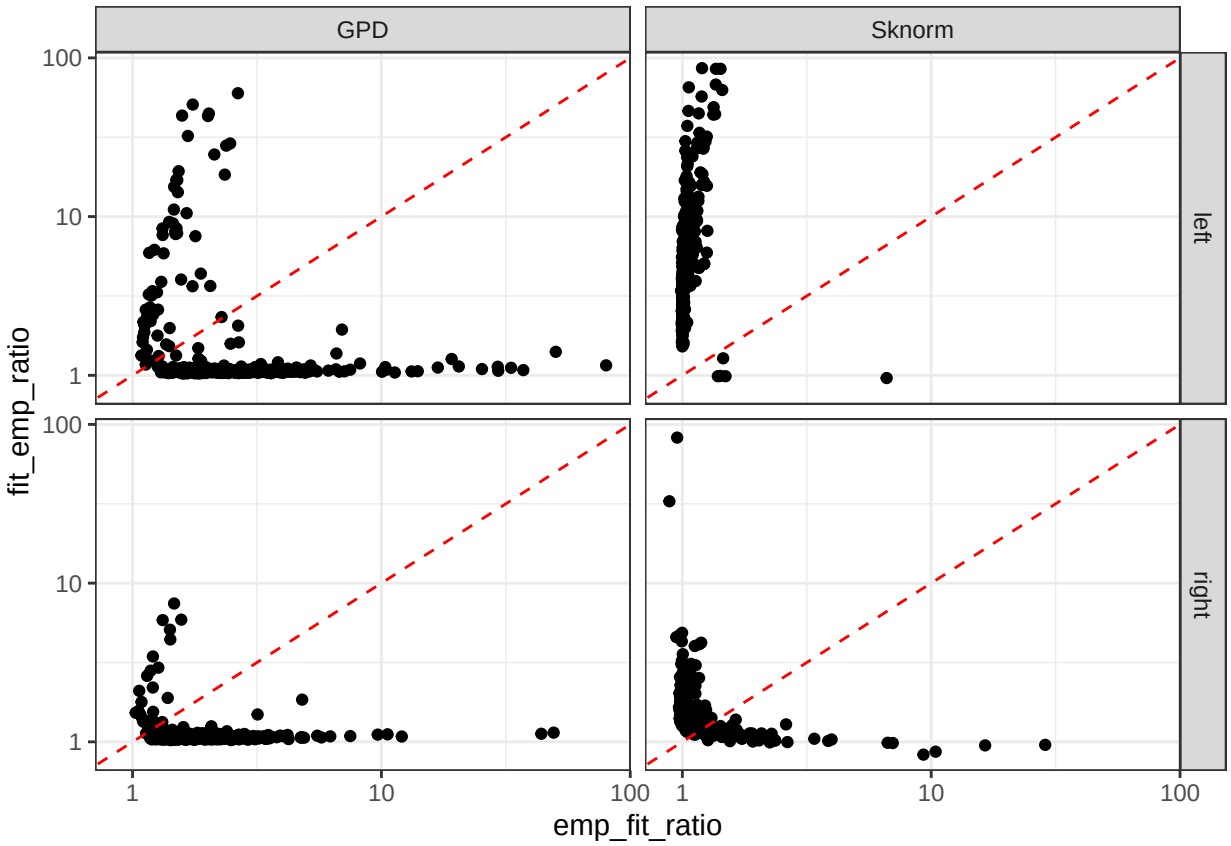
```
dot_plot <- param_comparison |>
  filter(EMP_GPD_ratio < 100 & EMP_SKN_ratio < 100) |>
  ggplot(aes_string(x = "EMP_GPD_ratio", y = "EMP_SKN_ratio")) +
  facet_grid(tail ~.) +
  geom_point() +
  geom_abline(linetype = "dashed", color = "red") +
  scale_x_log10() +
  scale_y_log10() +
  labs(title = "emp/fit p-value ratio plot")
```

```
hist_plot <- param |>
  filter(emp_fit_ratio < 100) |>
  ggplot(aes_string(x = "emp_fit_ratio")) +
  facet_grid(tail ~ method) +
  scale_x_log10() +
  geom_histogram() +
  geom_vline(xintercept = 10, linetype = "dashed", color = "red")
plot_grid(dot_plot,
          hist_plot,
          ncol = 2,
          align = "v")
```



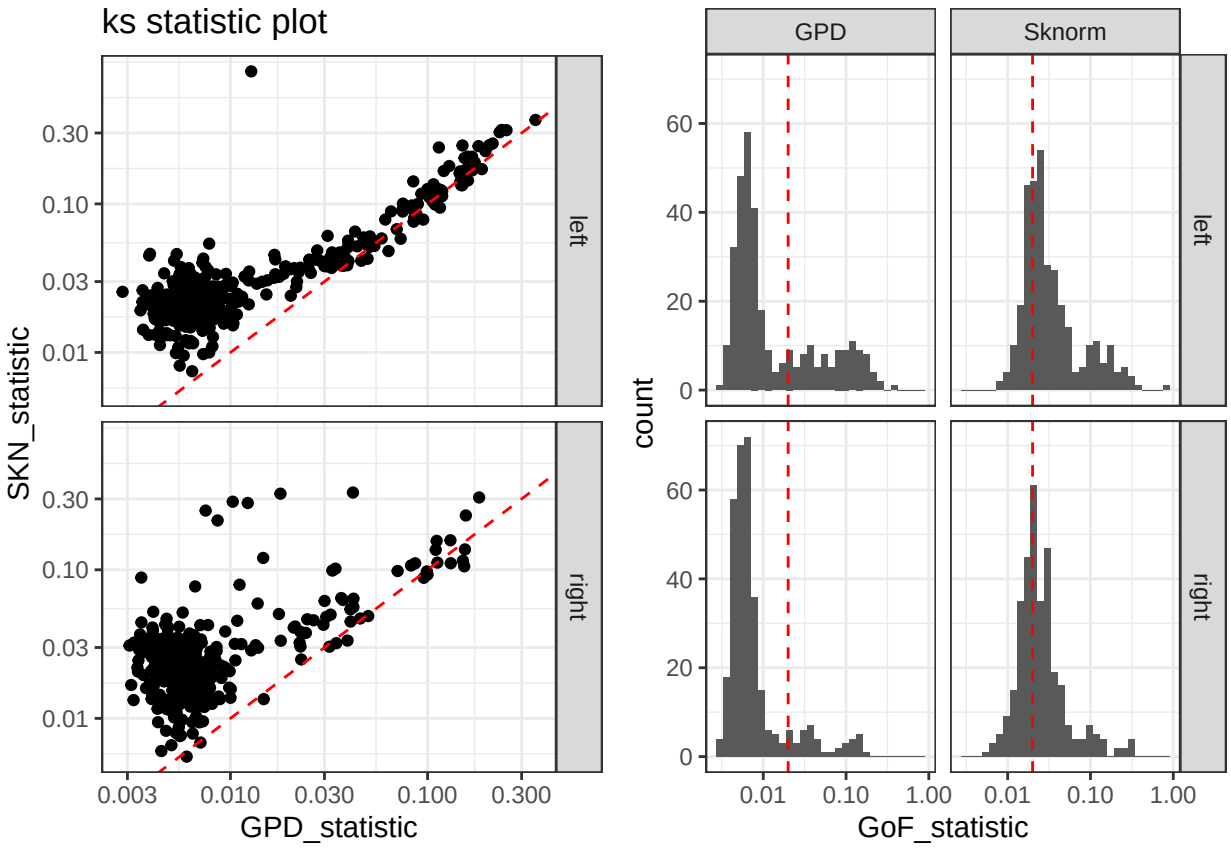
3. ratio compartion plot

```
dot_plot <- param |>
  filter(emp_fit_ratio < 100 & fit_emp_ratio < 100) |>
  ggplot(aes_string(x = "emp_fit_ratio", y = "fit_emp_ratio")) +
  facet_grid(tail ~ method) +
  scale_x_log10() +
  scale_y_log10() +
  geom_point() +
  geom_abline(linetype = "dashed", color = "red")
print(dot_plot)
```



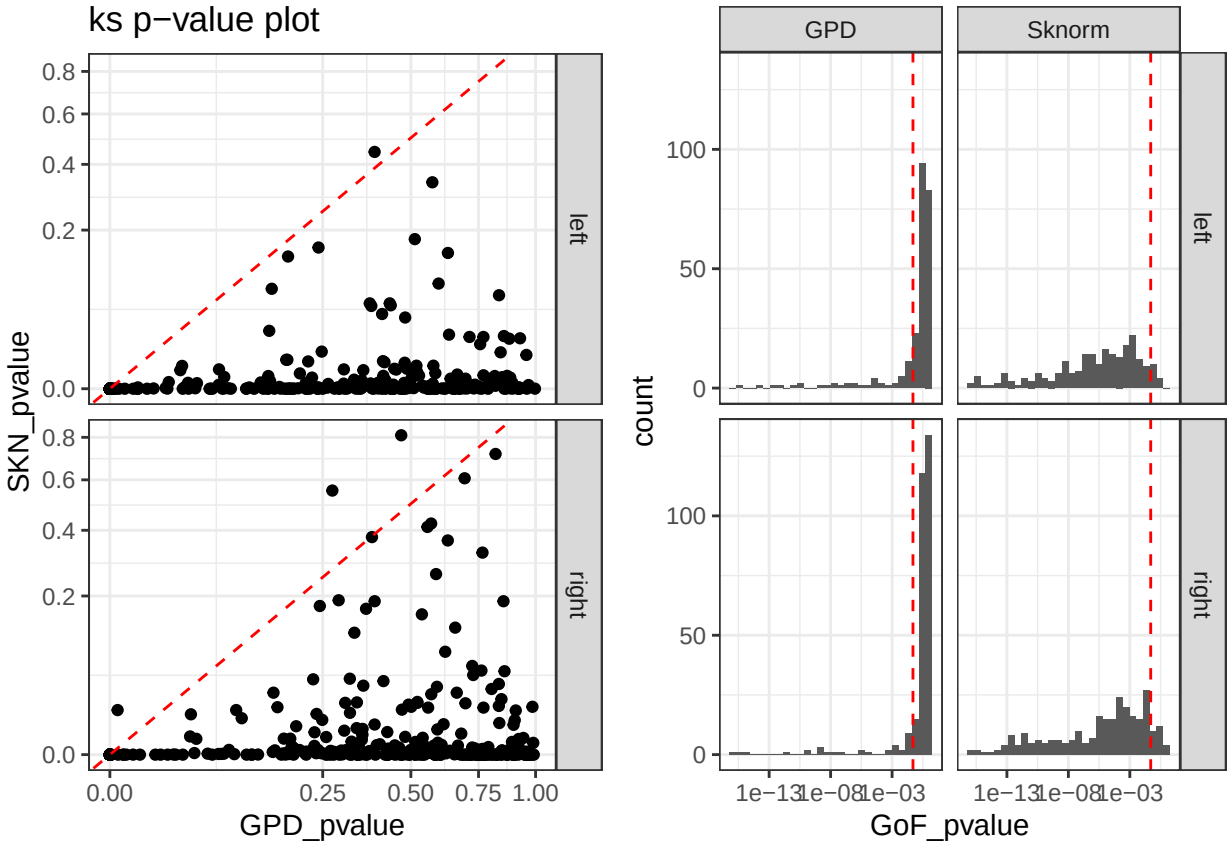
4. GoF statistic

```
dot_plot <- param_comparison |>
  ggplot(aes_string(x = "GPD_statistic", y = "SKN_statistic")) +
  facet_grid(tail ~.) +
  geom_point() +
  scale_x_log10() +
  scale_y_log10() +
  geom_abline(linetype = "dashed", color = "red") +
  labs(title = "ks statistic plot")

hist_plot <- param |>
  ggplot(aes_string(x = "GoF_statistic")) +
  facet_grid(tail ~ method) +
  scale_x_log10() +
  geom_histogram() +
  geom_vline(xintercept = 0.02, linetype = "dashed", color = "red")

plot_grid(dot_plot,
          hist_plot,
          ncol = 2,
          align = "v")
```



5. p-value of GoF test

```
dot_plot <- param_comparison |>
  ggplot(aes_string(x = "GPD_pvalue", y = "SKN_pvalue")) +
  facet_grid(tail ~.) +
  geom_point() +
  scale_x_sqrt() +
  scale_y_sqrt() +
  geom_abline(linetype = "dashed", color = "red") +
  labs(title = "ks p-value plot")

hist_plot <- param |>
  ggplot(aes_string(x = "GoF_pvalue")) +
  facet_grid(tail ~ method) +
  geom_histogram() +
  scale_x_log10() +
  geom_vline(xintercept = 0.05, linetype = "dashed", color = "red")

plot_grid(dot_plot,
          hist_plot,
          ncol = 2,
          align = "v")
```




```
# check if the inf value corresponds to bumpy distribution
# bumpy: 56, 143, 169, 175, 177, 185, 191, 198, 213, 236, 241, 243
# nonbumpy: 327, 325, 324
# both bumpy: 191
```

5. Summary

To summarize, let's take the median of each metric for both methods and both tails:

```
param |>
  group_by(method, tail) |>
  summarise(across(everything(), median), .groups = "drop") |>
  arrange(tail) |>
  rename(Method = method,
         Tail = tail,
         `KS statistic` = GoF_statistic,
         `KS p-value` = GoF_pvalue,
         `p-val overshoot factor` = fit_emp_ratio,
         `p-val undershoot factor` = emp_fit_ratio) |>
  mutate(`KS p-value` = num(`KS p-value`, notation = "sci"),
         `KS statistic` = num(`KS statistic`, notation = "sci"),
         Method = ifelse(Method == "Sknorm", "Skew-normal", Method)) |>
  kable(booktabs = TRUE, digits = 1) |>
  kable_styling(position = "center", latex_options = "hold_position")
```

Based on the KS test, we find that GPD is a much better fit than skew-normal. Based on the approximation of tail probabilities, a more mixed picture emerges. The GPD tends to *underestimate* the tail probability (by median factors of 2.4 and 1.8 in the left and right tails, respectively). On the other hand, skew-normal tends

Method	Tail	KS statistic	KS p-value	p-val overshoot factor	p-val undershoot factor
GPD	left	7.66e-3	1.91e-1	1.1	2.4
Skew-normal	left	2.62e-2	2.37e-8	4.3	1.0
GPD	right	6.14e-3	4.39e-1	1.1	1.8
Skew-normal	right	2.23e-2	3.35e-6	1.4	1.1

to *overestimate* the tail probability (by median factors of 4.3 and 1.4 in the left and right tails, respectively). The following table summarizes the pros and cons of GPD and skew-normal:

Property	Winner
Computation	GPD
Goodness of fit	GPD
Accommodates partially bumpy distributions	GPD
Overall closeness of fitted and empirical p-values	GPD
Has unbounded support	Skew-normal
Conservative p-value estimation	Skew-normal
Fit stability	Skew-normal
Fewer tuning parameters	Skew-normal