TCRemp distance properties

M.S.

2024-11-27

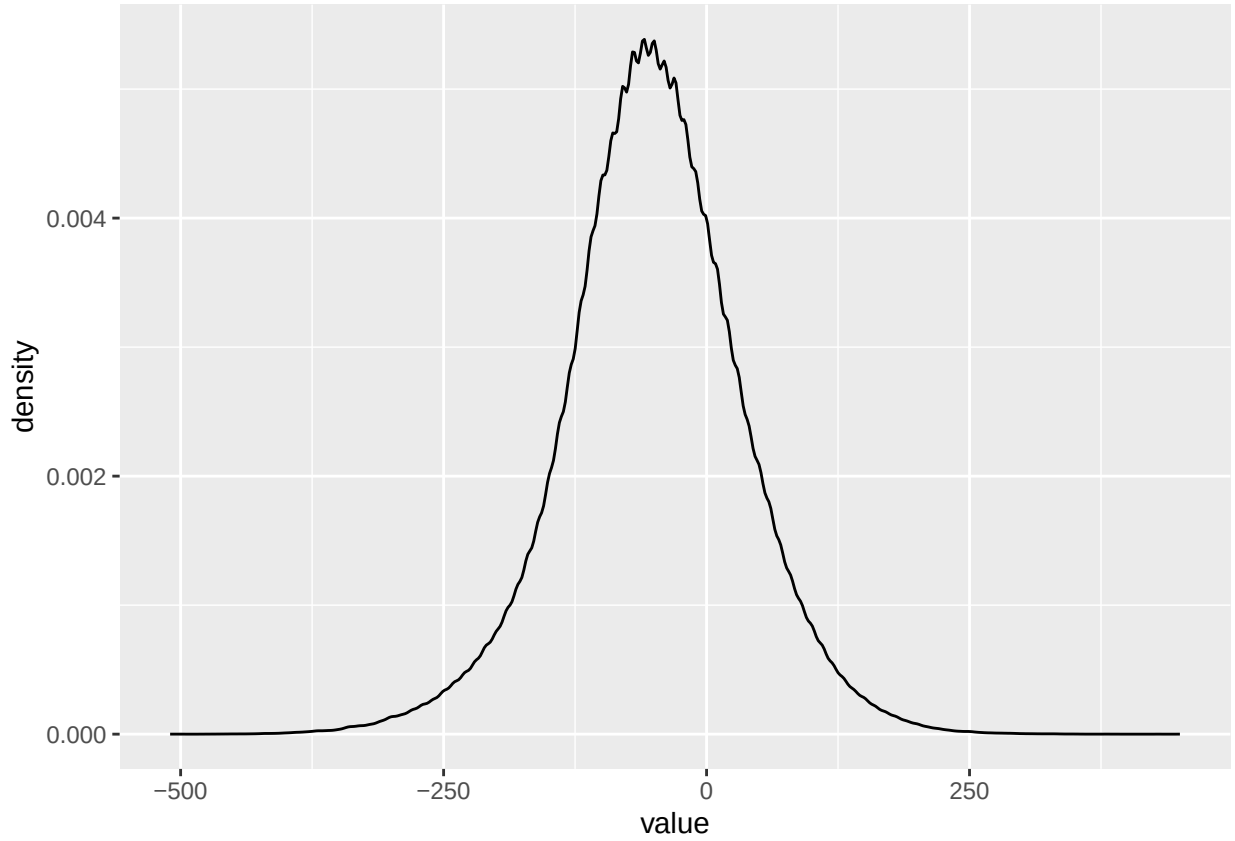
```
data <- read_tsv("res_TRB.txt.gz") |>
  rename(from = id)  |>
  mutate(from = as.character(from)) |>
  melt() |>
  filter(grepl("cdr3",variable))
```

```
## Rows: 994 Columns: 3001
## -- Column specification --------------------------------------------------------
## Delimiter: "\t"
## dbl (3001): id, 0_v_score, 0_j_score, 0_cdr3_score, 1_v_score, 1_j_score, 1_...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
## Using from as id variables
```

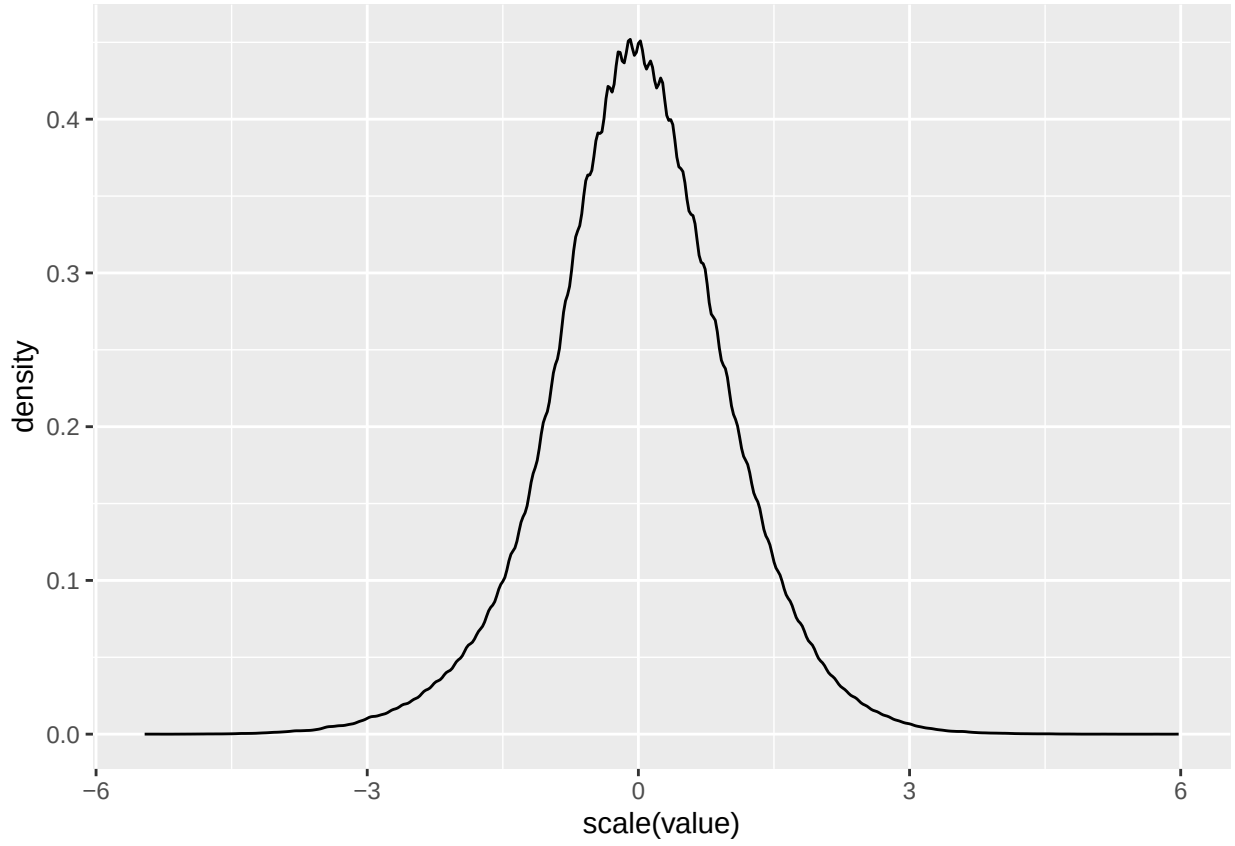
```
data$to <- str_split_fixed(data$variable, "_", 2)[,1]
data$variable <- NULL
data <- data |>
  mutate(from = paste0("x", from),
         to = paste0("x", to)) |>
  group_by(from) |>
  mutate(value.scaled = (value - mean(value)) / sd(value)) |>
  ungroup()
glimpse(data)
```

```
## Rows: 994,000
## Columns: 4
## $ from         <chr> "x0", "x1", "x2", "x3", "x4", "x5", "x6", "x7", "x8", "x9~
## $ value        <dbl> 730, -250, 150, -80, 10, 0, -250, -140, -230, -140, 10, 1~
## $ to           <chr> "x0", "x0", "x0", "x0", "x0", "x0", "x0", "x0", "x0", "x0~
## $ value.scaled <dbl> 9.8508303, -1.8381848, 2.2666303, -0.7458501, 0.9840220, ~
```

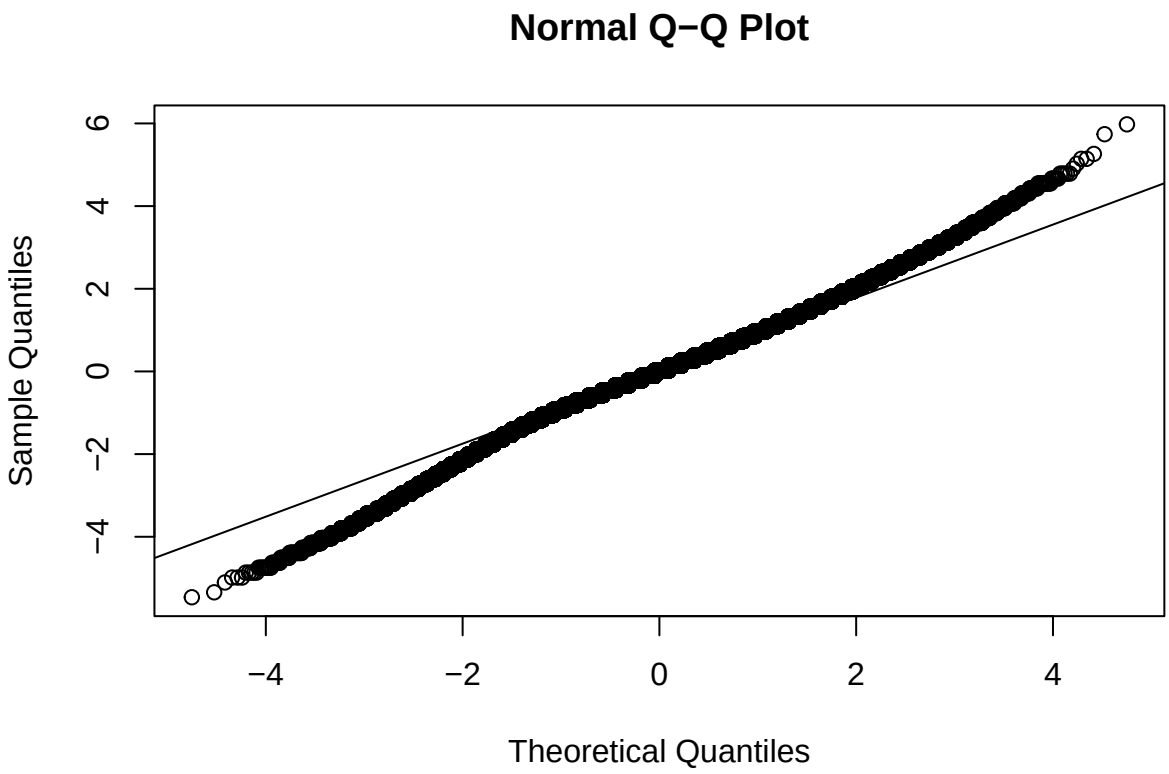
```
data1 <- data |> filter(as.character(to) > as.character(from))

ggplot(data1, aes(x = value)) +
  geom_density()
```




```
ggplot(data1, aes(x = scale(value))) +
  geom_density()
```




```
qqnorm(scale(data1$value))
qqline(scale(data1$value))
```



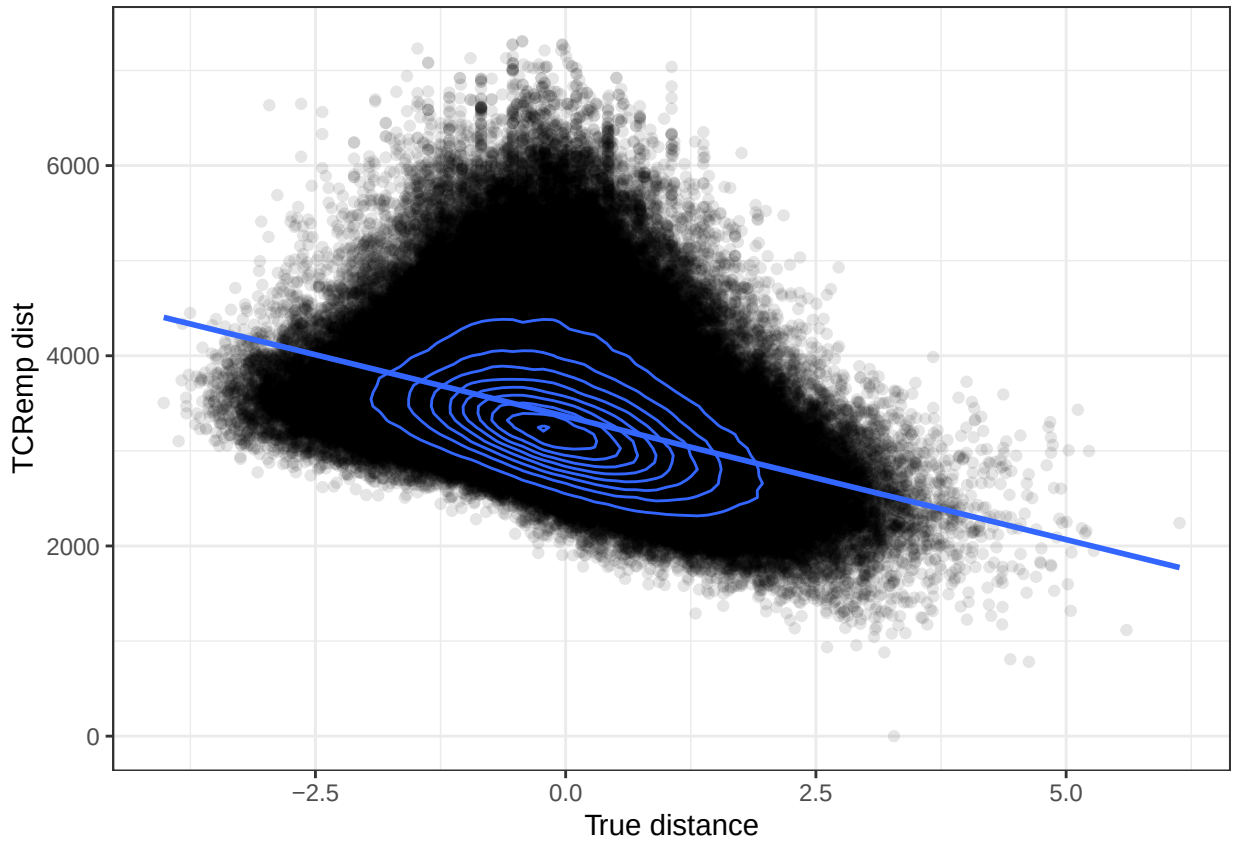

```
data.m <- data |>
  dcast(from ~ to)
rownames(data.m) <- data.m$from
data.m$from <- NULL
data.m <- as.matrix(data.m)
```

```
dists <- dist(data.m) |>
  as.matrix() |>
  melt() |>
  rename(from = Var1, to = Var2, dist_eucl = value) |>
  mutate(from = as.character(from), to = as.character(to))
```

```
data.comb <- data1 |>
  merge(dists)
```

```
data.comb |>
  ggplot(aes(x = value.scaled,
             y = dist_eucl)) +
  geom_point(alpha = 0.1) +
  geom_density_2d() +
  geom_smooth(method = "lm") +
  xlab("True distance") +
  ylab("TCRemp dist") +
  theme_bw()
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
cor.test(data.comb$value.scaled, data.comb$dist_eucl, method = "pearson")
```

```
##
##  Pearson's product-moment correlation
##
## data:  data.comb$value.scaled and data.comb$dist_eucl
## t = -320.65, df = 493519, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4175348 -0.4129169
## sample estimates:
##        cor
## -0.4152285
```

```
cor.test(data.comb$value.scaled, data.comb$dist_eucl, method = "spearman")
```

```
## Warning in cor.test.default(data.comb$value.scaled, data.comb$dist_eucl, :
## Cannot compute exact p-value with ties
```

```
##
##  Spearman's rank correlation rho
##
## data:  data.comb$value.scaled and data.comb$dist_eucl
## S = 2.9619e+16, p-value < 2.2e-16
```

```
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## -0.478468
```

```
#FIN
```