Stability After Batch Correction: 1 PC vs 2 PCs

Tecla Duran Fort

2025-09-05

Contents

Load Data

```
df <- read.csv("../../data/peak_table_var.csv")

# Peak table matrix (samples x peaks)
X <- as.matrix(df %>% dplyr::select(starts_with("Cluster")))
stopifnot(nrow(X) == nrow(df))

# Batch encoding
batch <- df$batch
nbatch <- length(unique(batch))
B <- model.matrix(~ 0 + factor(batch))
colnames(B) <- paste0("Batch_", sort(unique(batch)))
```

Correcting Only Batch

Setup

```
# Batch-wise means (samples -> batches)
batch_sizes <- as.vector(t(B) %*% rep(1, nrow(X)))
batch_sums  <- t(B) %*% X
M           <- sweep(batch_sums, 1, batch_sizes, "/")  # (nbatch x nvars)

# Map means back to sample space (scores live in sample space)
X_batch_means <- B %*% M

# PCA on batch-wise mean intensities
pca_batch <- prcomp(X_batch_means, scale. = TRUE)
```

```
pc_scores <- pca_batch$x            # (samples x PCs) -> sample-space PCs
pc_axis   <- axis_labels_from_pca(pca_batch)
```

Build 1-PC and 2-PC Batch Corrections

```
# 1 PC removal: project out PC1 from X
X_batch_1PC <- orthogonal_correction(X, pc_scores[,1])$corrected

# 2 PCs removal: sequentially remove PC1 then PC2
X_batch_2PC <- orthogonal_correction(orthogonal_correction(X, pc_scores[,1])$corrected, pc_scores[,2])$
```

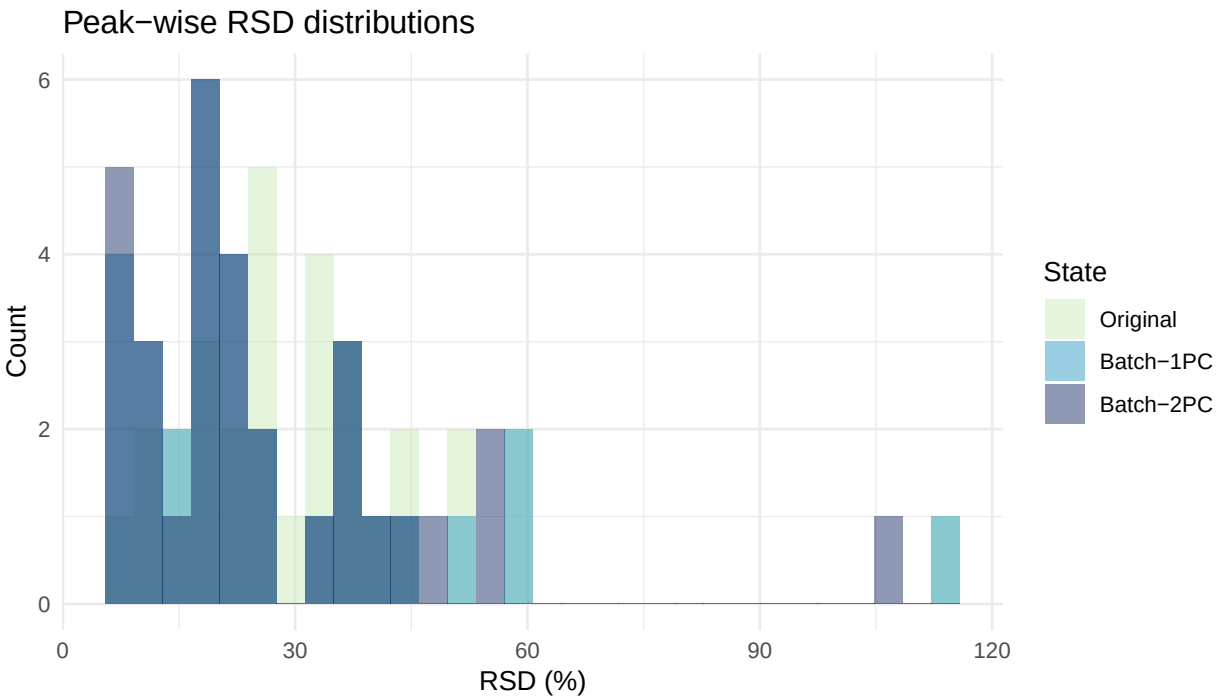
Results

Stability (RSD) Before vs After

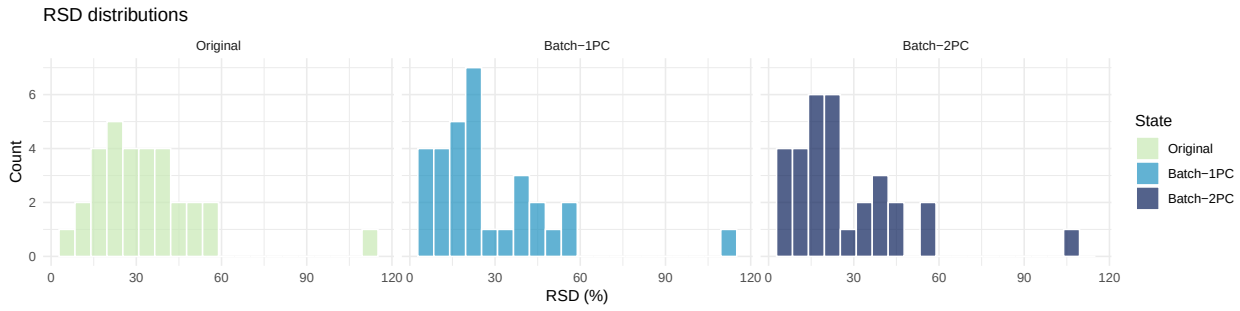
```
# RSD per peak (column-wise)
rsd_before   <- apply(X,            2, rsd_vec)
rsd_after_1  <- apply(X_batch_1PC,  2, rsd_vec)
rsd_after_2  <- apply(X_batch_2PC,  2, rsd_vec)

rsd_long <- tibble(
  peak = colnames(X),
  Original = rsd_before,
  `Batch-1PC` = rsd_after_1,
  `Batch-2PC` = rsd_after_2
) |>
  pivot_longer(-peak, names_to = "State", values_to = "RSD") |>
  filter(!is.na(RSD))
rsd_long$State <- factor(rsd_long$State,
                         levels = c("Original", "Batch-1PC", "Batch-2PC"))
# Quick summary table
rsd_summary <- rsd_long |>
  group_by(State) |>
  summarize(
    n = n(),
    median_RSD = median(RSD),
    mean_RSD = mean(RSD),
    pct_RSD_lt_20 = mean(RSD < 20)*100,
    .groups = "drop"
  )

kable(rsd_summary, digits = 2, caption = "RSD summary by state") |>
  kable_styling(full_width = FALSE, position = "center")
```

Table 1: RSD summary by state

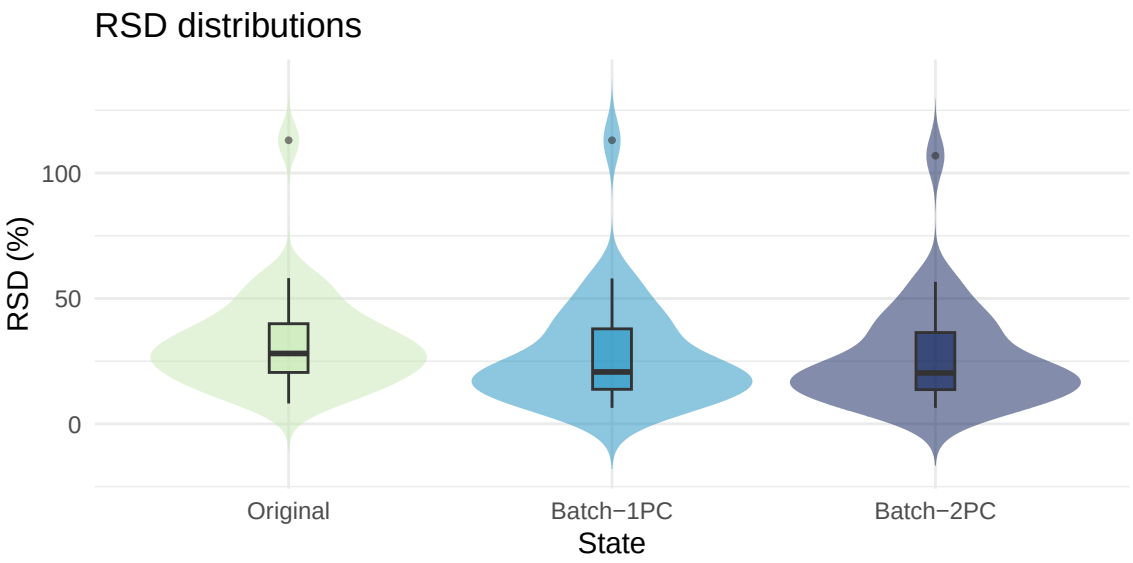
State	n	median_RSD	mean_RSD	pct_RSD_lt_20
Original	31	28.07	32.75	22.58
Batch-1PC	31	20.70	27.39	41.94
Batch-2PC	31	20.31	26.55	45.16



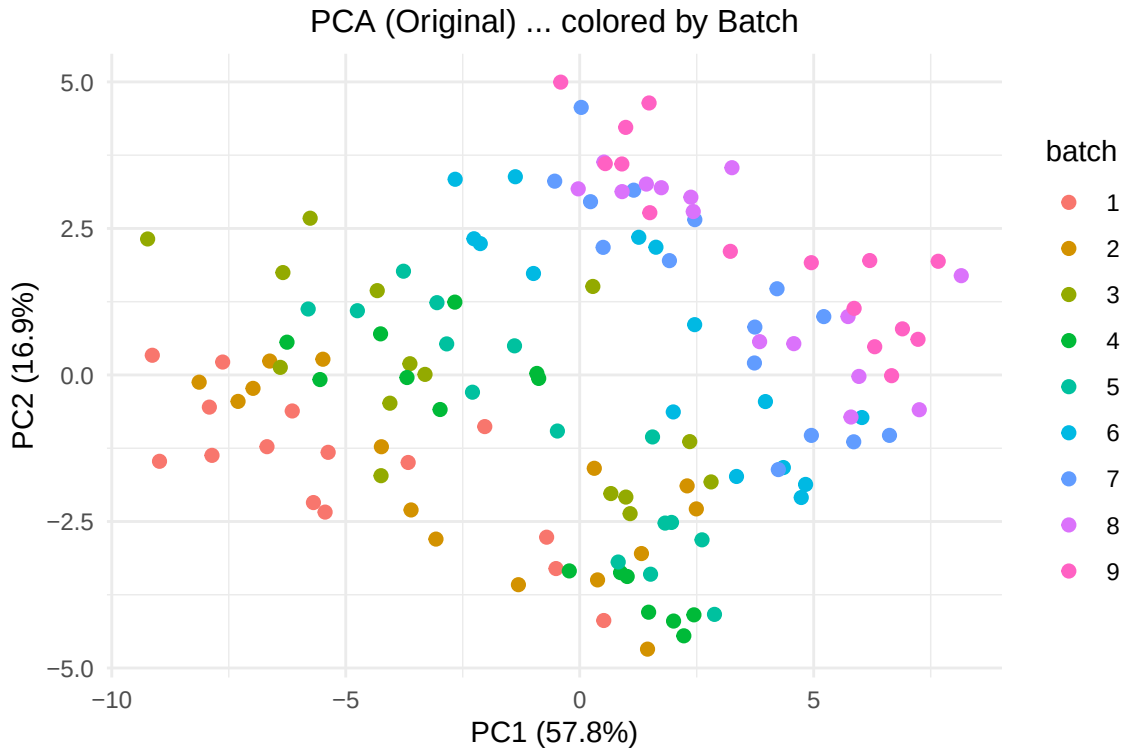
Histograms

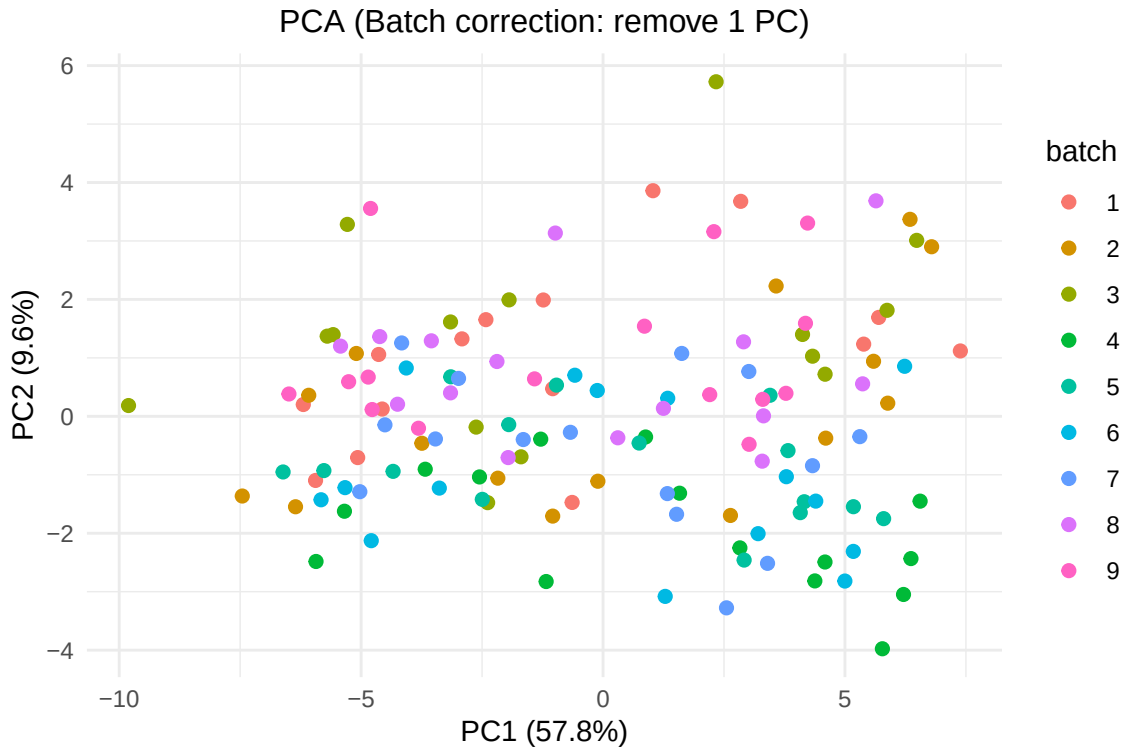


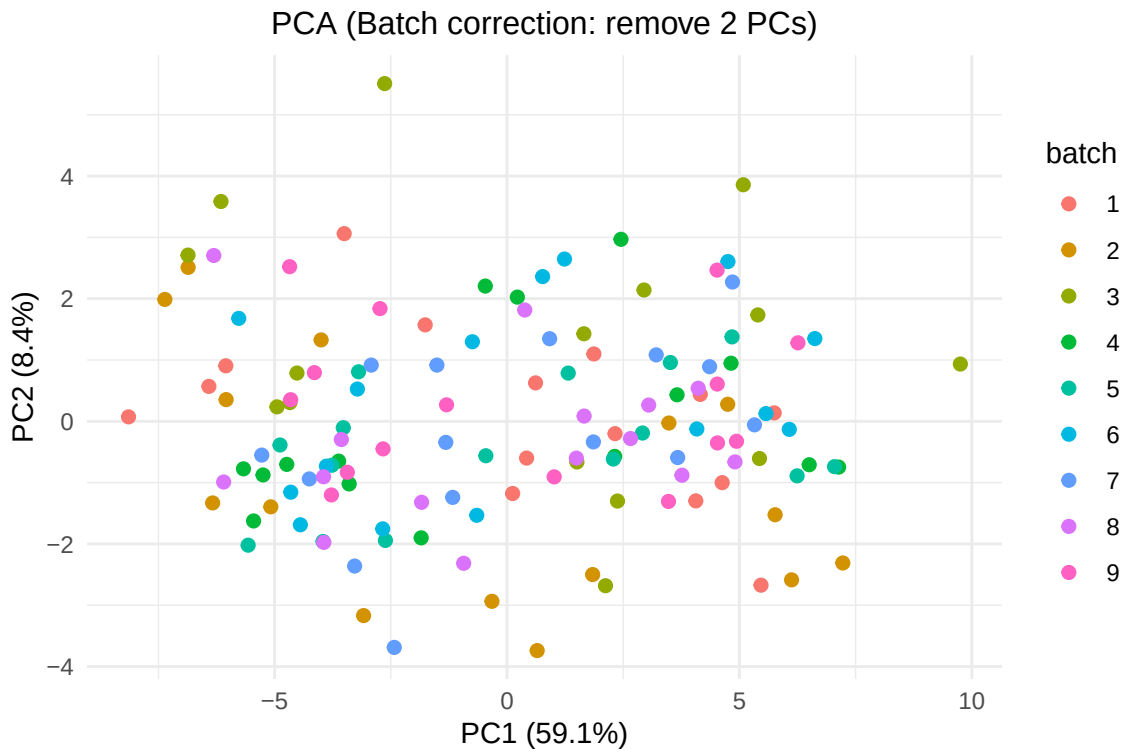
Violin plots .



PCA Colored by Batch







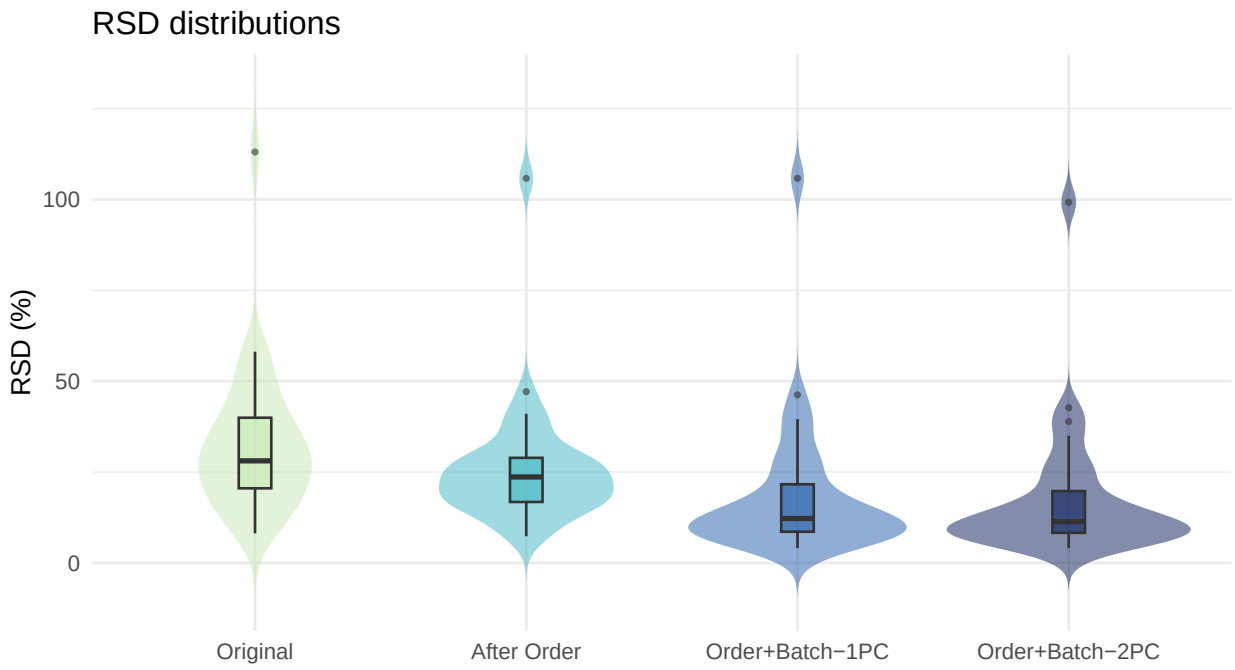
Results Correcting Order (time) First

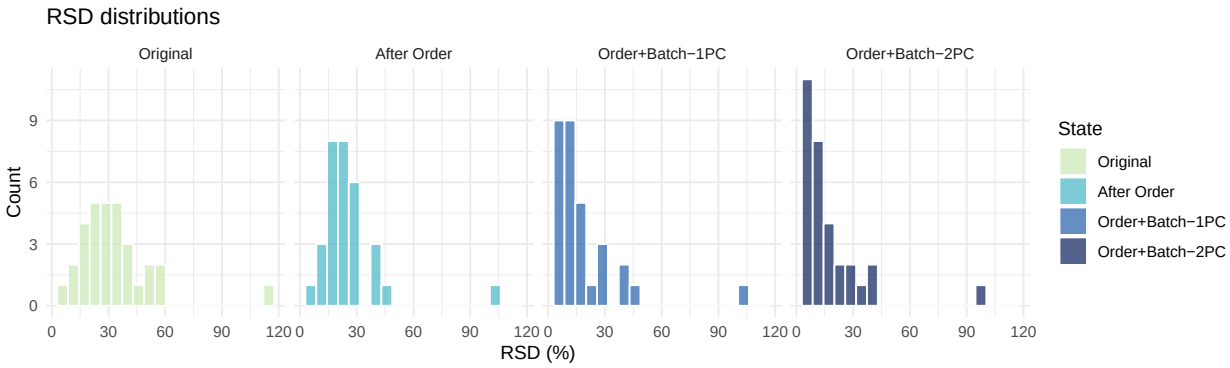
```r
order_in_batch <- df %>%
  arrange(batch, elapsed_time) %>%
  group_by(batch) %>%
  mutate(order_in_batch = row_number()) %>%
  pull(order_in_batch)

# One-hot d'ordre
max_order <- max(order_in_batch)
O <- model.matrix(~ 0 + factor(order_in_batch, levels = 1:max_order))

# Mitjanes per ordre i PCA per obtenir PC1 en espai de mostres
order_sizes <- as.vector(t(O) %*% rep(1, nrow(X)))
order_sums  <- t(O) %*% X
M_order     <- sweep(order_sums, 1, order_sizes, "/")
X_order_means <- O %*% M_order

pca_order <- prcomp(X_order_means, scale. = TRUE)
pc_order  <- pca_order$x[, 1]

# Correcció per ordre
X_order_corr <- orthogonal_correction(X, pc_order)$corrected

# Batch PCs (sobre mitjanes de batch després de corregir temps)
B <- model.matrix(~ 0 + factor(batch))
batch_sizes <- as.vector(t(B) %*% rep(1, nrow(X)))
batch_sums  <- t(B) %*% X_order_corr
```

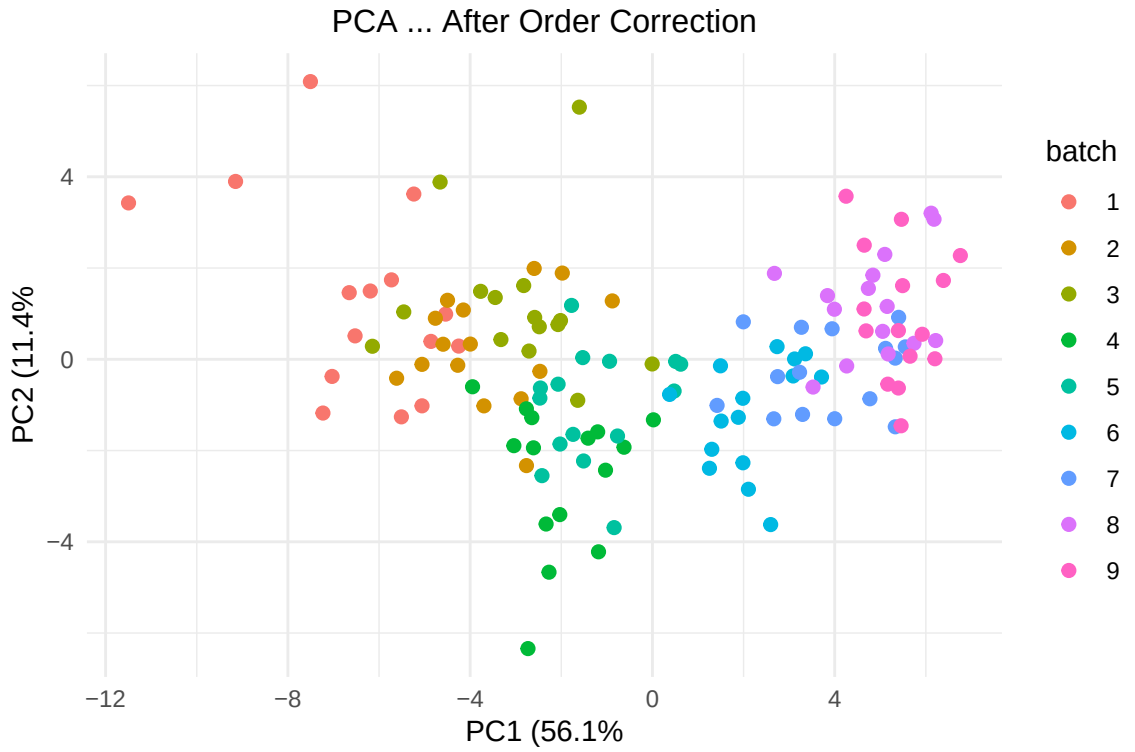
```r
M_batch     <- sweep(batch_sums, 1, batch_sizes, "/")
X_batch_means <- B %*% M_batch
pca_batch <- prcomp(X_batch_means, scale. = TRUE)
pc_scores <- pca_batch$x

# Versions 1PC i 2PC
X_order_batch_1PC <- orthogonal_correction(X_order_corr, pc_scores[,1])$corrected
X_order_batch_2PC <- orthogonal_correction(
  orthogonal_correction(X_order_corr, pc_scores[,1])$corrected,
  pc_scores[,2])$corrected
```

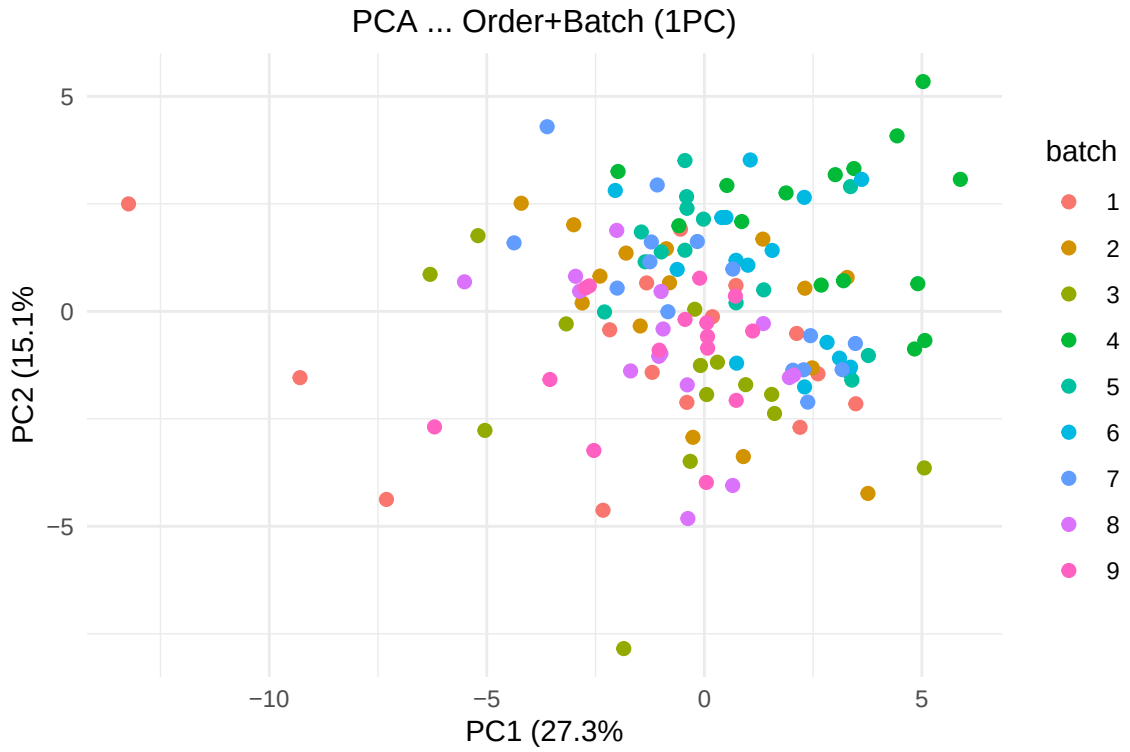
Table 2: RSD summary (Original, After Order, Order+Batch-1PC, Order+Batch-2PC)

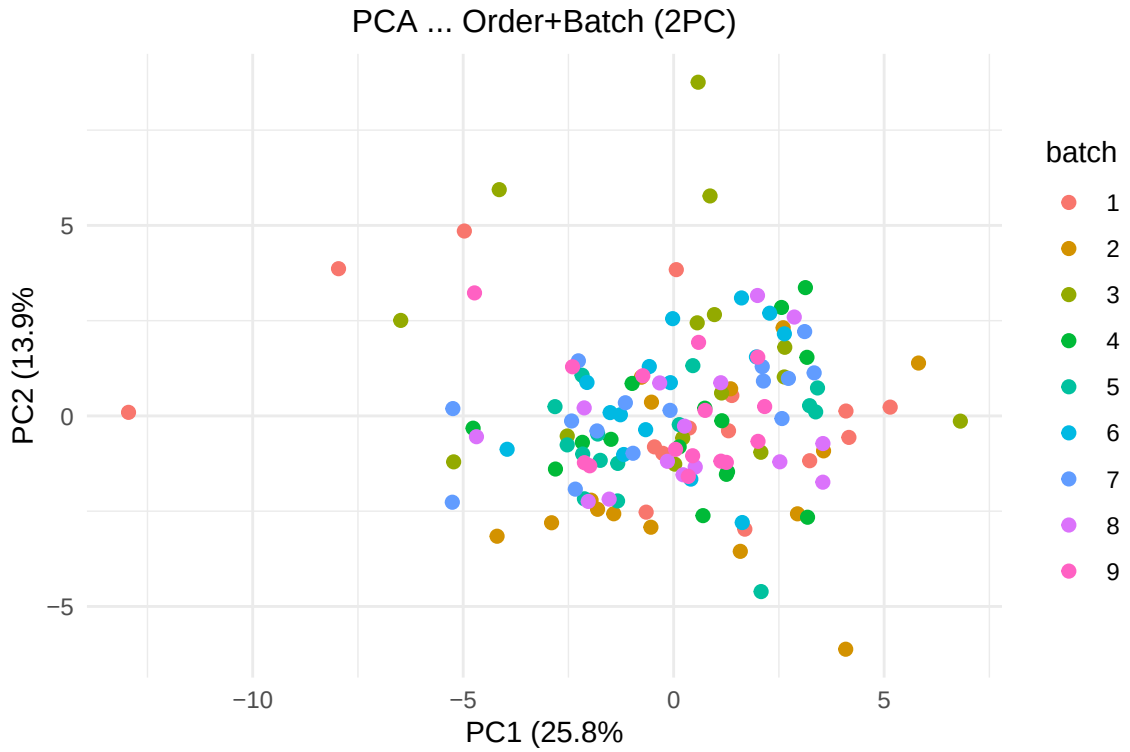
State	n	median_RSD	mean_RSD	pct_under_20
Original	31	28.07	32.75	22.58
After Order	31	23.64	26.14	38.71
Order+Batch-1PC	31	12.21	18.80	74.19
Order+Batch-2PC	31	11.34	17.66	74.19




RSD distributions
Original
After Order
Order+Batch−1PC
Order+Batch−2PC
Count
RSD (%)
State
Original
After Order
Order+Batch−1PC
Order+Batch−2PC


PCA ... After Order Correction
PC2 (11.4%)
PC1 (56.1%
batch
1
2
3
4
5
6
7
8
9


PCA ... Order+Batch (1PC)
PC2 (15.1%)
PC1 (27.3%
batch
1
2
3
4
5
6
7
8
9
5
0
−5
−10
−5
0
5


PCA ... Order+Batch (2PC)
PC2 (13.9%)
PC1 (25.8%
batch
1
2
3
4
5
6
7
8
9
5
0
−5
−10
−5
0
5