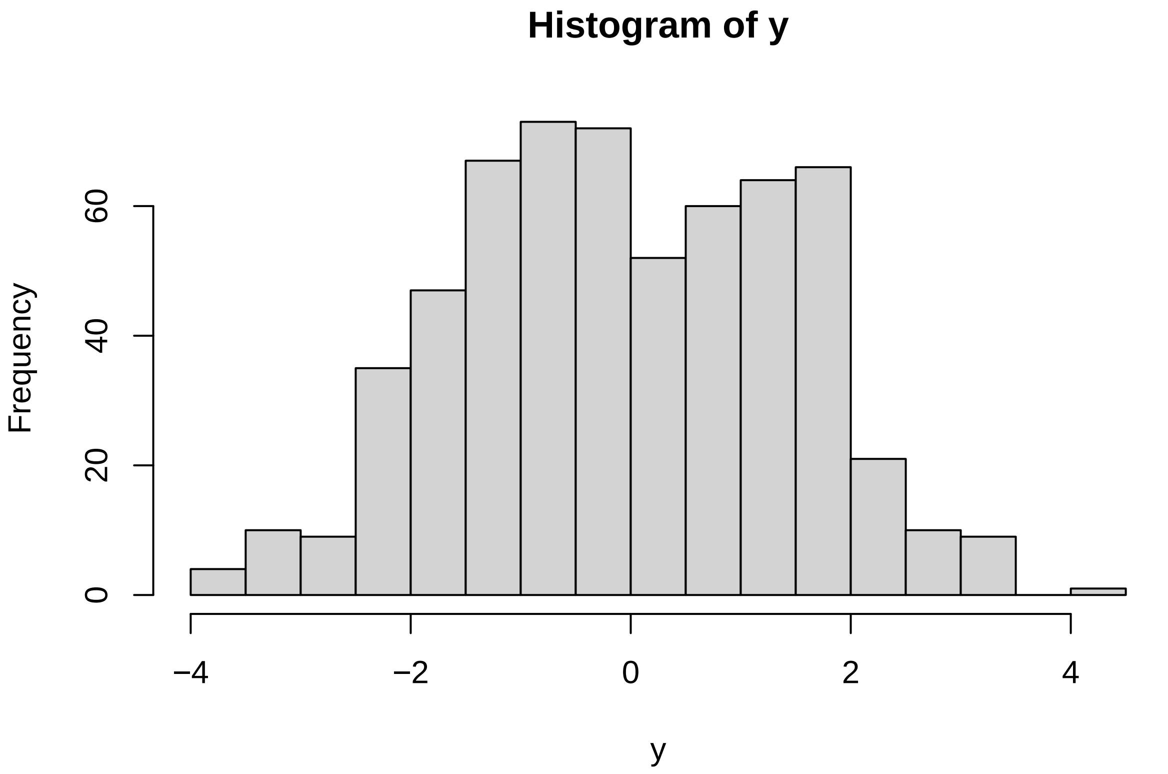
Binder

Guanyu

2025-08-10

```
devtools::load_all()
library(WASABI.ext)
library(BNPmix)
library(lpSolve)
library(mcclust)
library(salso)
library(superheat)
library(ggplot2)
```

```
set.seed(12345)
mu <-  c(-1.1, 1.1)
prop <- c(0.5, 0.5)
n <- 600
components <- sample(1:2, size = n, replace = TRUE, prob = prop)
y <- rnorm(n, mean = mu[components], sd = 1)
hist(y, breaks = 20)
```

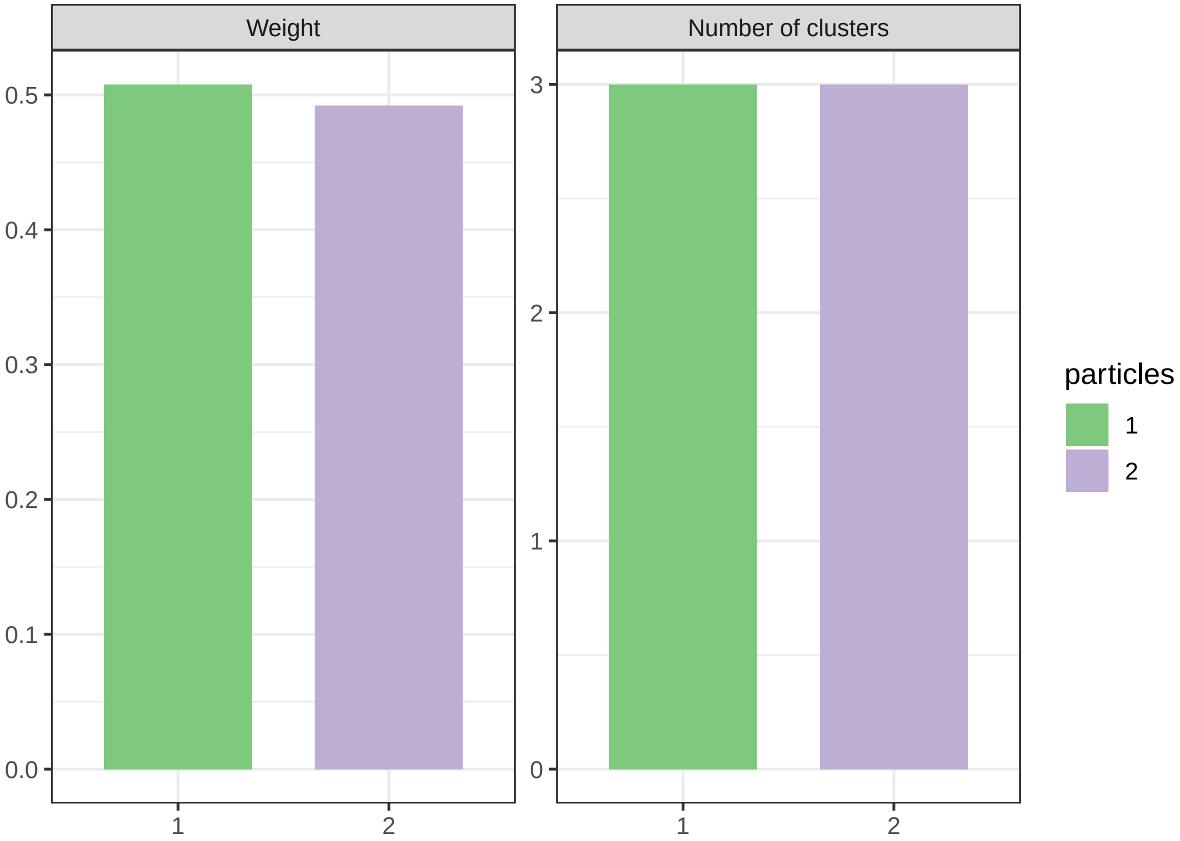


```
est_model <- BNPmix::PYdensity(y = y,
                               mcmc = list(niter = 15000,
                                           nburn = 5000,
                                           model = "LS",
                                           print_message = FALSE),
                               output = list(out_type = "FULL",
                                             out_param = TRUE))
cls.draw = est_model$clust
z_minb <- salso::salso(cls.draw, loss = binder(a = 1.3))
table(z_minb)
```

```
## z_minb
##   1   2
## 375 225
```

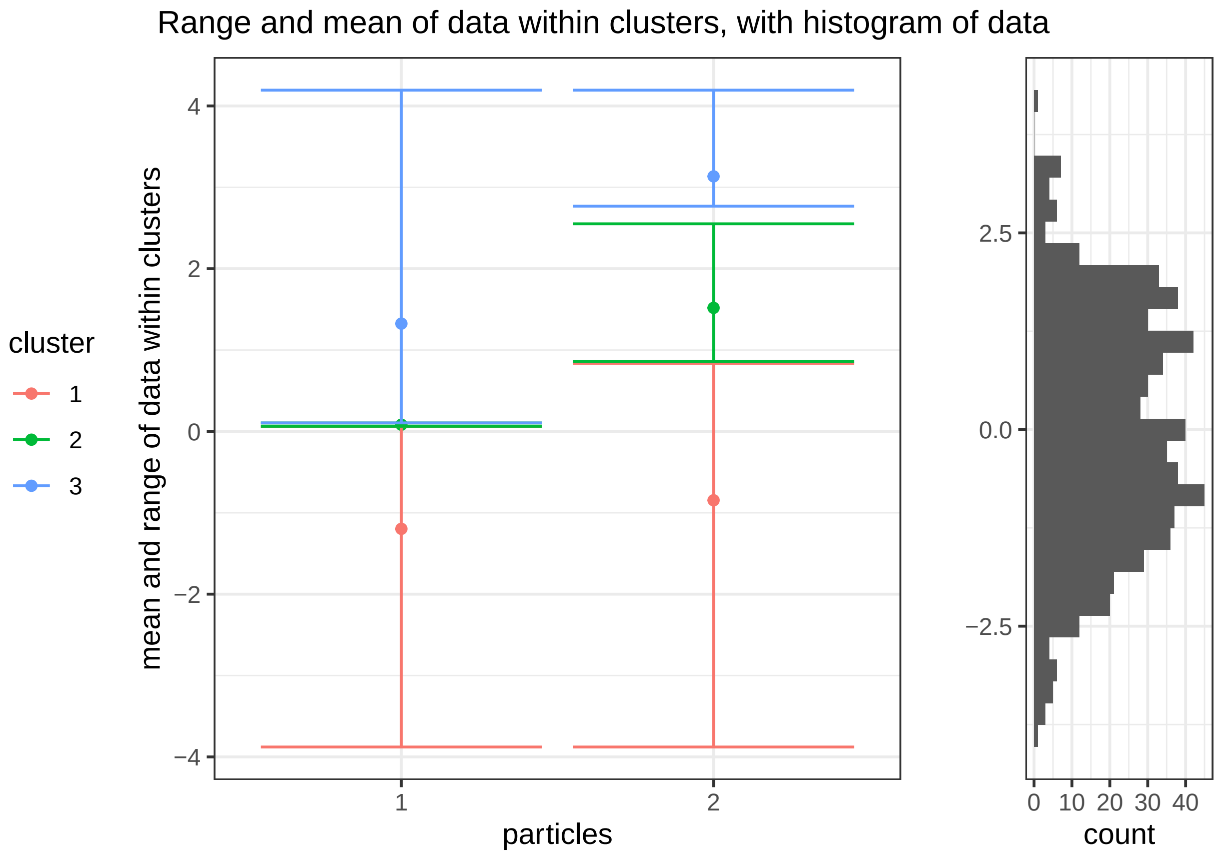
```
psm=mcclust::comp.psm(cls.draw+1)
```

```
out_WASABI <- WASABI(cls.draw, psm = psm, L = 2,
                     method.init = "topvi", method = "salso",
                     loss = "Binder", a = 1.3, maxNClusters = 10)

out_WASABI_ms <- WASABI_multistart(cls.draw, psm = psm, L = 2,
                                   multi.start = 20, ncores = 4,
                                   mini.batch = 150,
                                   max.iter = 10, extra.iter = 4,
                                   method.init = "++", method = "salso",
                                   a = 1.2,
                                   loss = "Binder", maxNClusters = 10)

if(out_WASABI_ms$wass.dist < out_WASABI$wass.dist){
  out_WASABI <- out_WASABI_ms
}
ggsummary(out_WASABI)
```

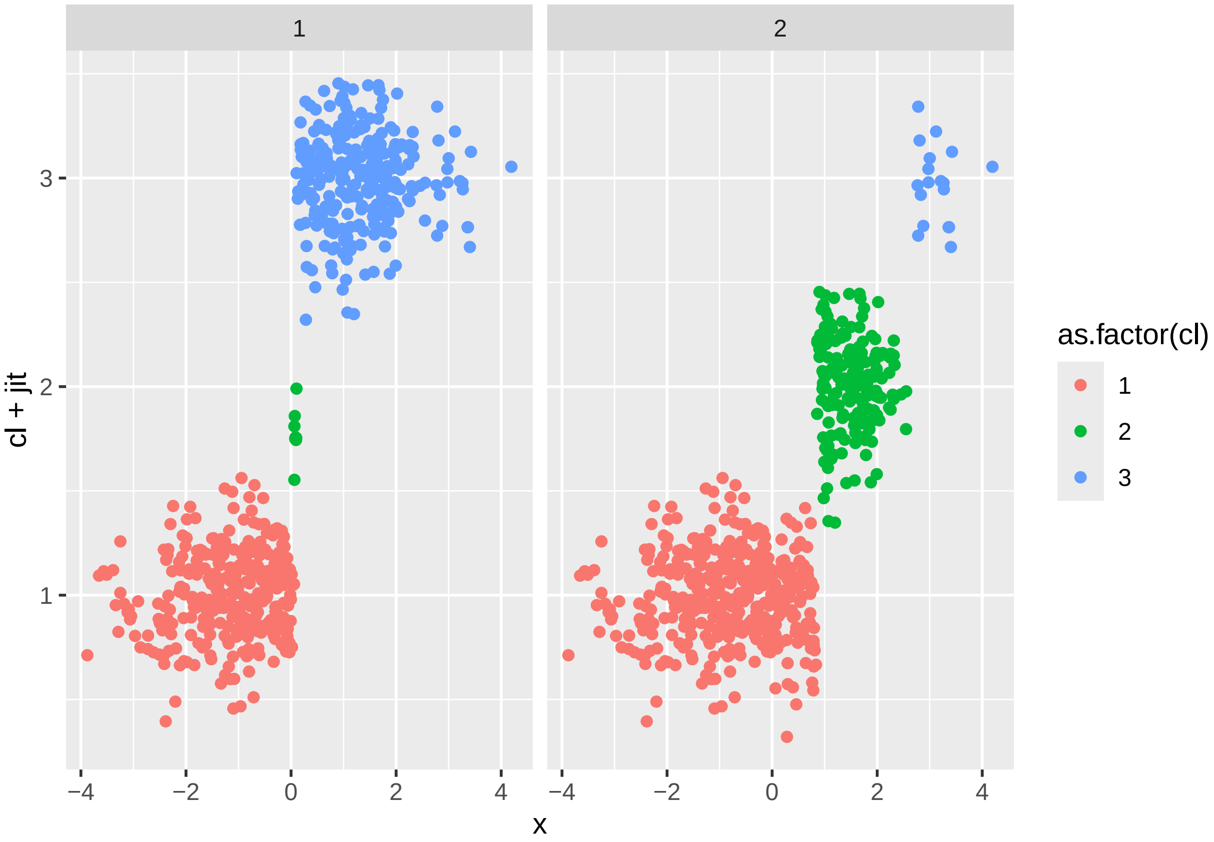


```
ggrange_hist(out_WASABI, y)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



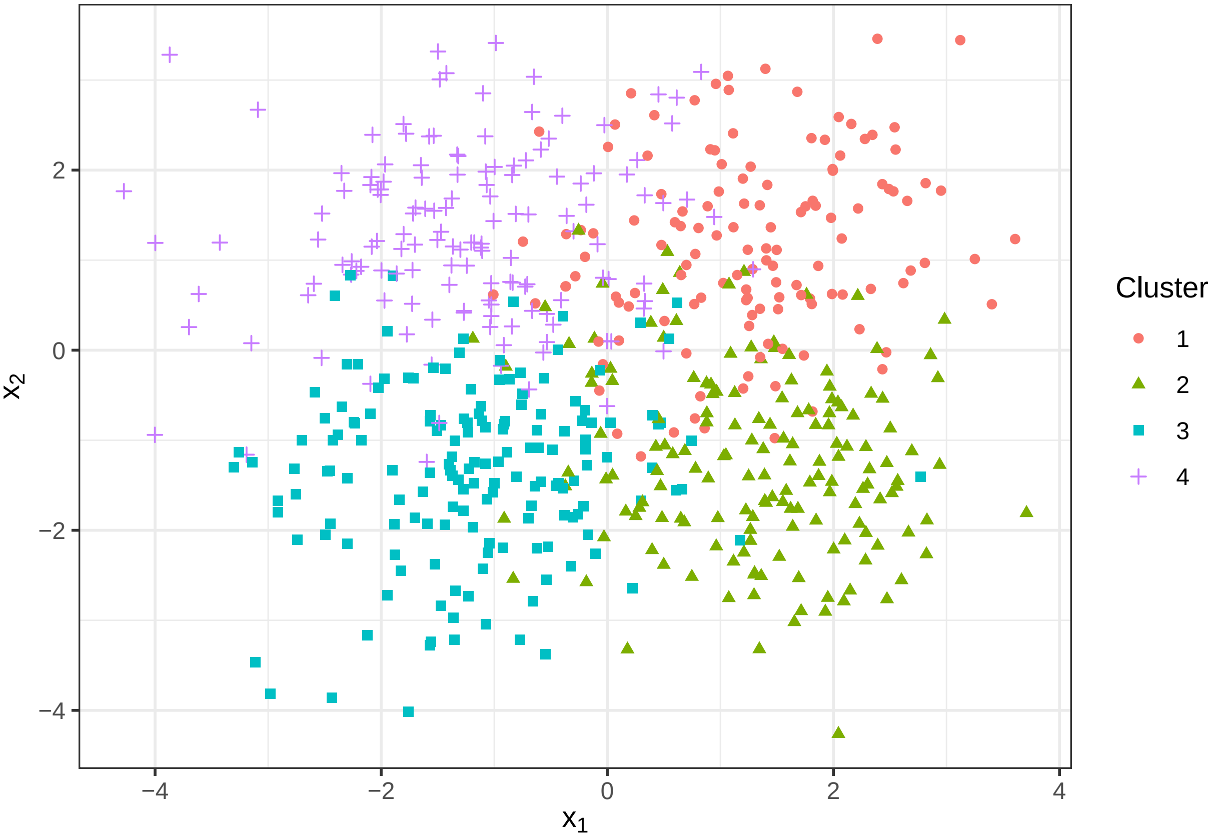
```
ggscatter_grid(out_WASABI, y)
```



Two-dimensional data

```
m = 1.25
n = 600
p = 2
Kt = 4

set.seed(4321)

Y=matrix(rnorm(p*n),n,p)
usim=runif(n)
ind=ifelse(usim<1/4,1,ifelse(usim<1/2,2,ifelse(usim<3/4,3,4)))
Y[ind==1,] = Y[ind==1,] +m
Y[ind==2,1] = Y[ind==2,1] + m; Y[ind==2,2] = Y[ind==2,2] - m;
Y[ind==3,] = Y[ind==3,] -m
Y[ind==4,1] = Y[ind==4,1] - m; Y[ind==4,2] = Y[ind==4,2] + m;

cls.true = ind
```

```
library(ggplot2)
ggplot() +
  geom_point(aes(x = Y[,1],
                 y = Y[,2],
                 colour = as.factor(cls.true),
                 shape  = as.factor(cls.true))) +
  theme_bw() + guides(colour=guide_legend(title="Cluster"),
                    shape = guide_legend(title="Cluster")) +
  xlab(expression("x"[1])) + ylab(expression("x"[2]))
```



```
set.seed(4321)
### Parameters for DP mixture
alpha = 1
# using Fraley and Raftery recommendation
a_x=rep((p+2)/2,p)
khat = 4
b_x= rep(mean(apply(Y,2,var))/(khat^(2/p))/2,p)

### Parameters for MCMC function
S=10000 # 10000
thin = 1
tot = S*thin
burnin= 5000 # 5000

est_model <- BNPmix::PYdensity(y = Y,
                       mcmc = list(niter = burnin + tot,
                                   nburn = burnin,
                                   model = "DLS",
                                   hyper = FALSE
                                   ),
                       prior = list(
                         k0 = 0.1*rep(1,p),
                         a0 = a_x,
                         b0 = b_x,
                         strength = alpha,
                         discount = 0),
                       output = list(out_type = "FULL", out_param = TRUE))
```

```
## Completed:   1500/15000 - in 0.527026 sec
## Completed:   3000/15000 - in 1.01599 sec
## Completed:   4500/15000 - in 1.49361 sec
## Completed:   6000/15000 - in 2.12113 sec
## Completed:   7500/15000 - in 2.7361 sec
## Completed:   9000/15000 - in 3.42296 sec
## Completed:   10500/15000 - in 4.1259 sec
## Completed:   12000/15000 - in 4.78954 sec
## Completed:   13500/15000 - in 5.43034 sec
## Completed:   15000/15000 - in 6.09335 sec
##
## Estimation done in 6.09341 seconds
```

```
cls.draw = est_model$clust
psm=mcclust::comp.psm(cls.draw+1)
```

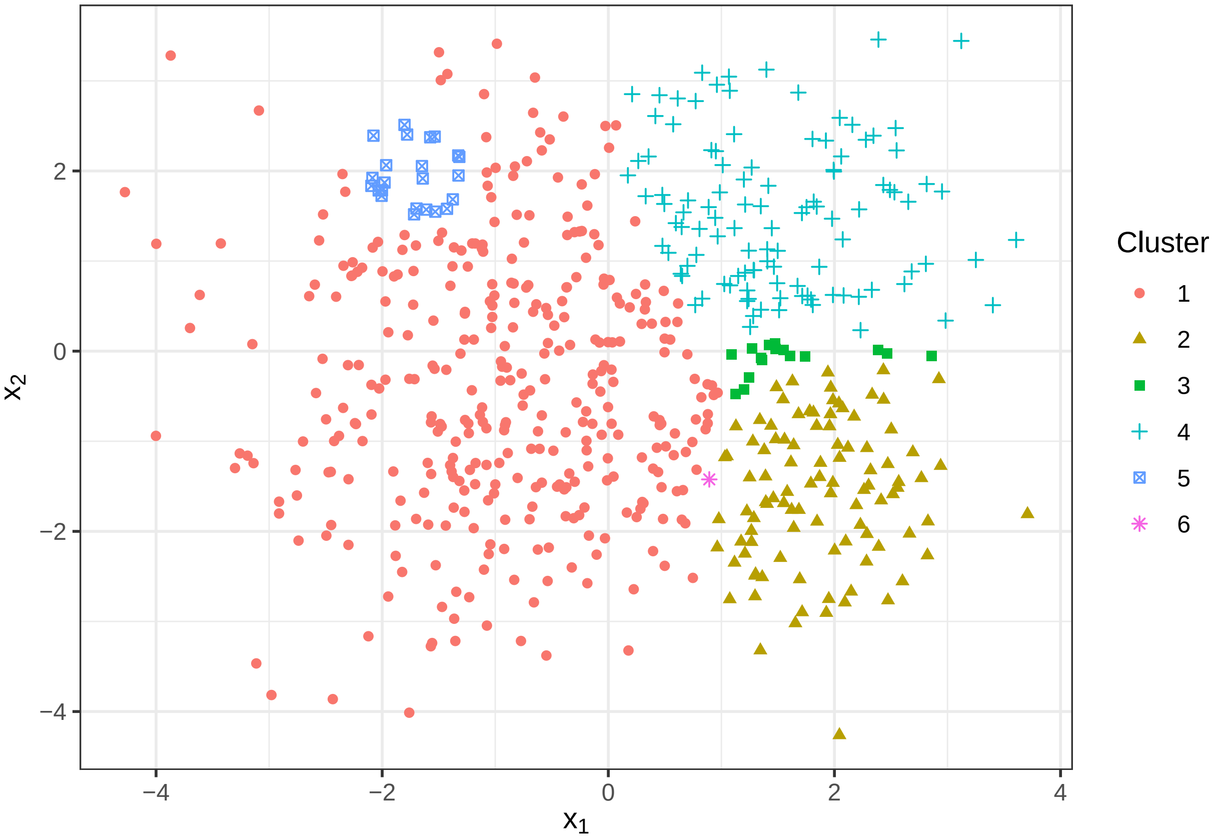
The following shows how WASABI works for different value of ‘a’.

a = 1.1

```
z_minb <- salso::salso(cls.draw, loss = binder(a = 1.1), maxNClusters = 10)
table(z_minb)
```

```
## z_minb
##   1   2   3   4   5   6
## 351  99  16 110  23   1
```

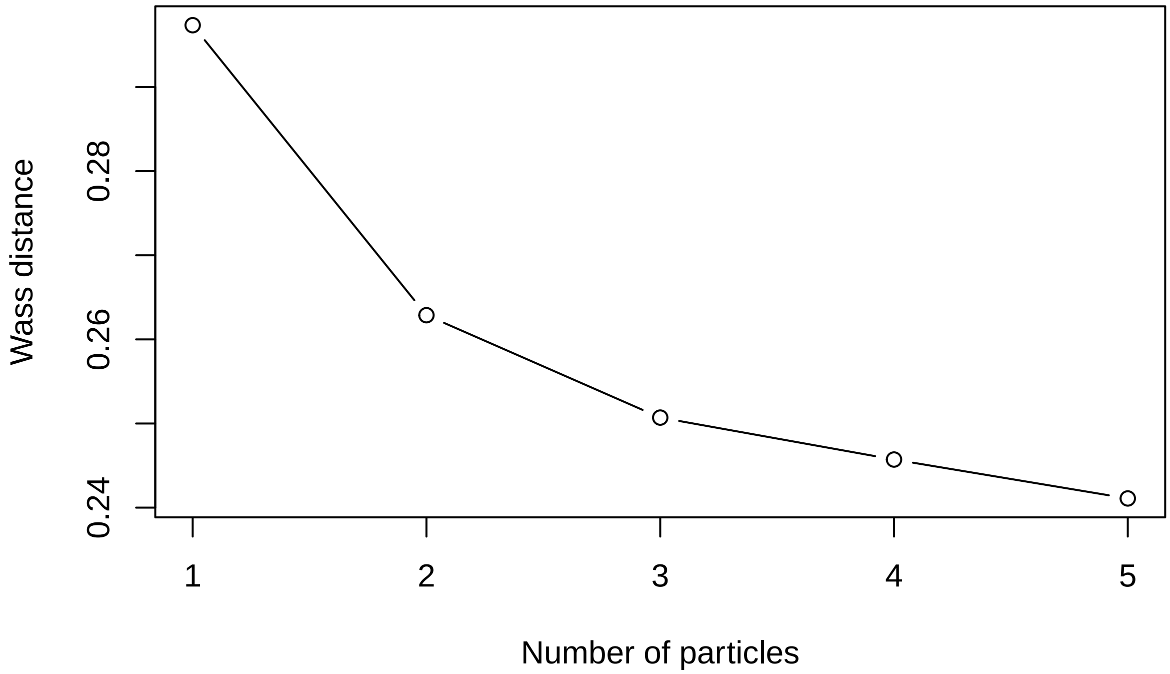
```
df = data.frame(x1 = Y[,1],
                x2 = Y[,2],
                Cluster = z_minb)
df$Cluster = as.factor(df$Cluster)

ggplot(df)+
  geom_point(aes(x = x1, y = x2, color = Cluster, shape = Cluster)) +
  ylab(expression("x"[2]))+xlab(expression("x"[1]))+
  theme_bw()
```



```r
set.seed(123)
out_elbow <- elbow(cls.draw, L_max = 5, psm = psm,
                   multi.start = 1,
                   method.init = "++", method = "salso",
                   loss = "Binder", a = 1.1, maxNClusters = 10)
```

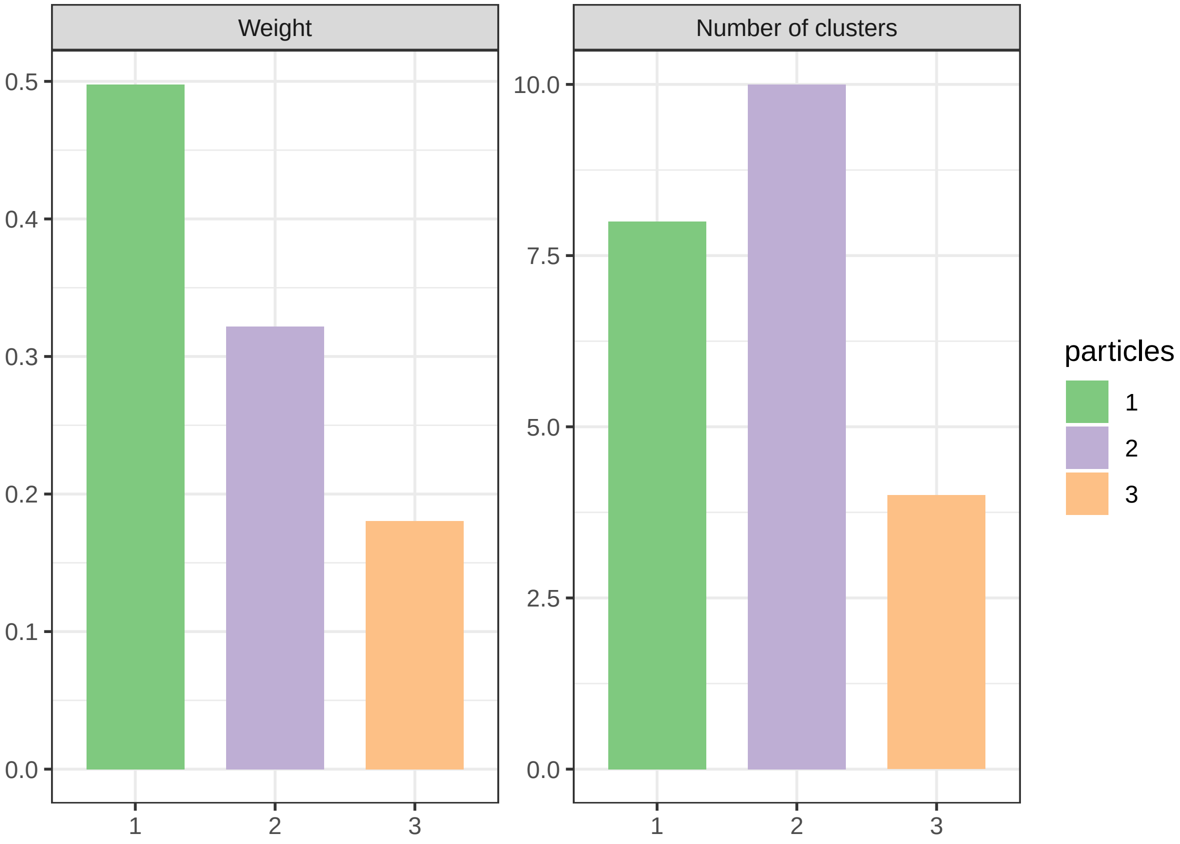
```
## Completed  1 / 5
## Completed  2 / 5
## Completed  3 / 5
## Completed  4 / 5
## Completed  5 / 5
```

```r
plot(out_elbow$wass_vec, type = "b", ylab = "Wass distance", xlab = "Number of particles")
```

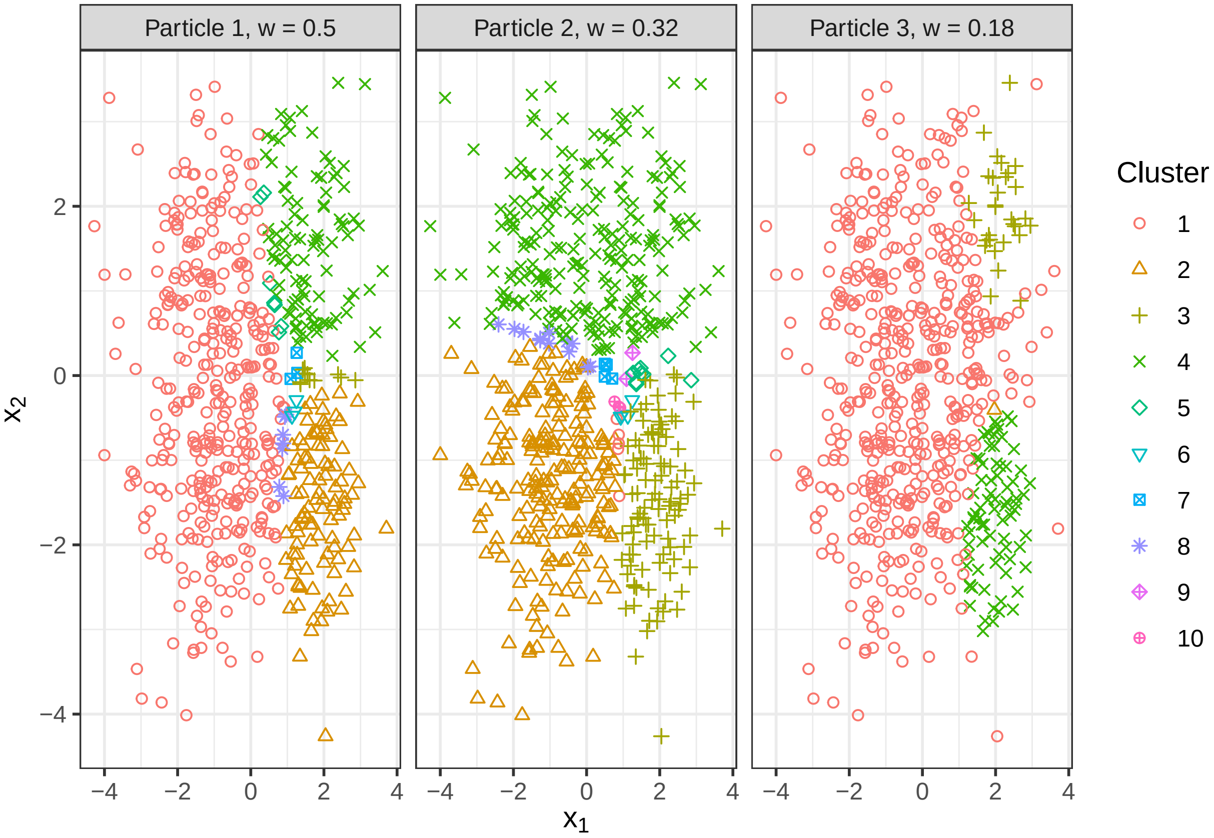


We choose “L=3” as the optimal number of clusters.

```
L = 3
output_WASABI <- out_elbow$output_list[[L]]
# output_WASABI_mb = WASABI_multistart(cls.draw, psm,
#                                      multi.start = 25, ncores = 4,
#                                      method.init ="++", add_topvi = FALSE,
#                                      method="salso", L=L,
#                                      mini.batch = 500,
#                                      max.iter= 10, extra.iter = 5,
#                                      suppress.comment=TRUE,
#                                      swap_countone = TRUE,
#                                      seed = 54321, loss = "Binder",
#                                      a = 1.1,
#                                      maxNClusters = 10)
#
# if(output_WASABI_mb$wass.dist < output_WASABI$wass.dist){
#   output_WASABI <- output_WASABI_mb
# }

ggsummary(output_WASABI)
```



```
ggscatter_grid2d(output_WASABI, Y)
```

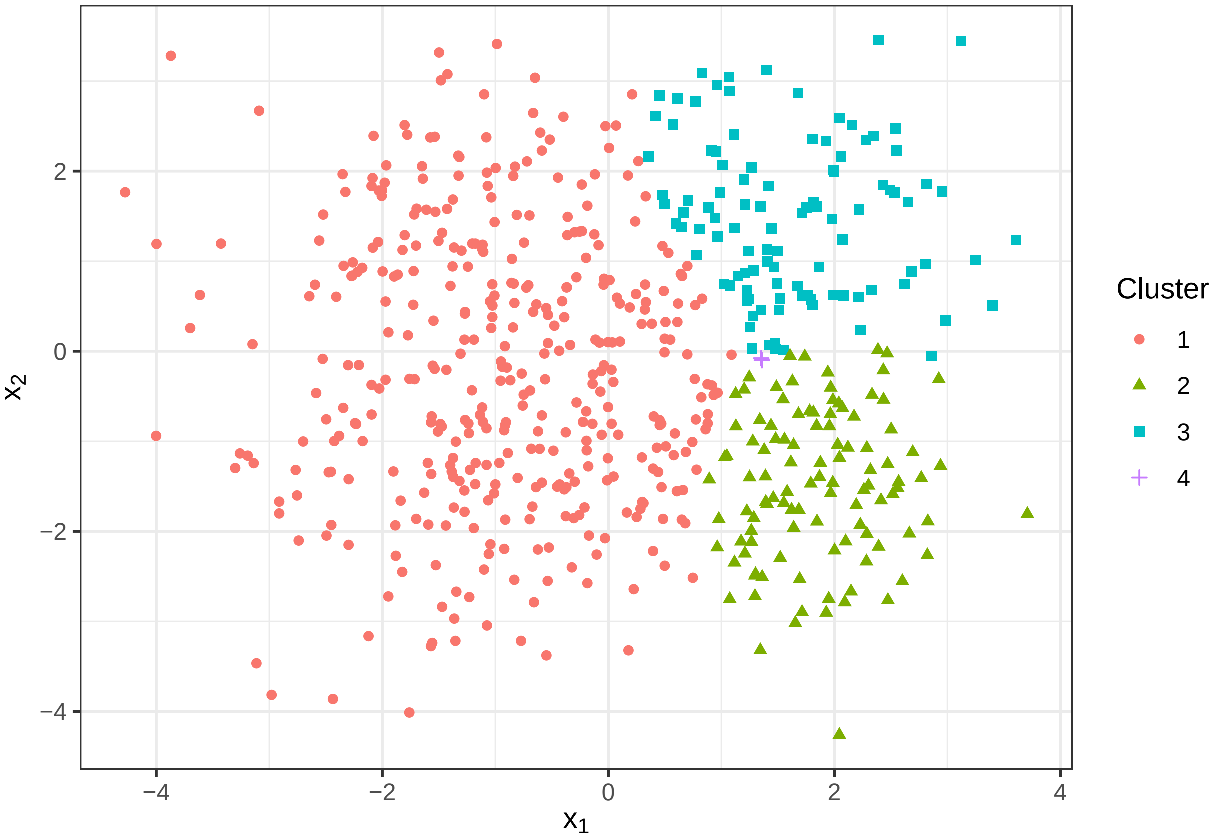


a = 1.2

```
z_minb <- salso::salso(cls.draw, loss = binder(a = 1.2), maxNClusters = 10)
table(z_minb)
```

```
## z_minb
##   1   2   3   4
## 386 107 105   2
```

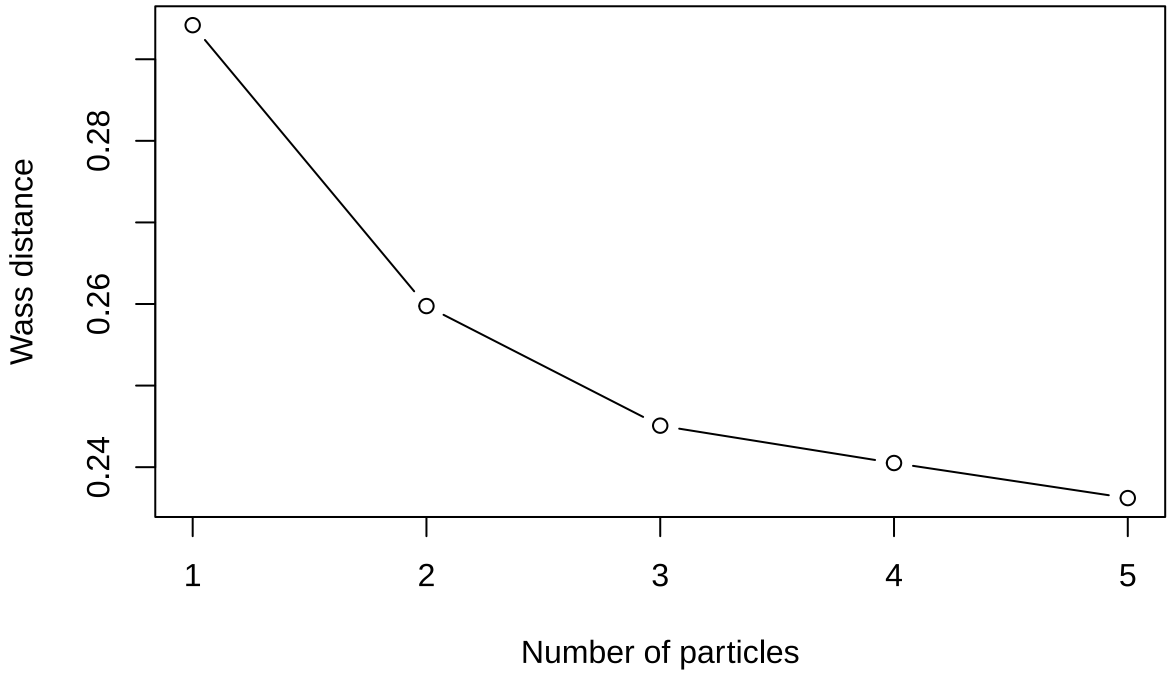
```
df = data.frame(x1 = Y[,1],
                x2 = Y[,2],
                Cluster = z_minb)
df$Cluster = as.factor(df$Cluster)

ggplot(df)+
  geom_point(aes(x = x1, y = x2, color = Cluster, shape = Cluster)) +
  ylab(expression("x"[2]))+xlab(expression("x"[1]))+
  theme_bw()
```



```
set.seed(123)
out_elbow <- elbow(cls.draw, L_max = 5, psm = psm,
                   multi.start = 1,
                   method.init = "++", method = "salso",
                   loss = "Binder", a = 1.2, maxNClusters = 10)
```

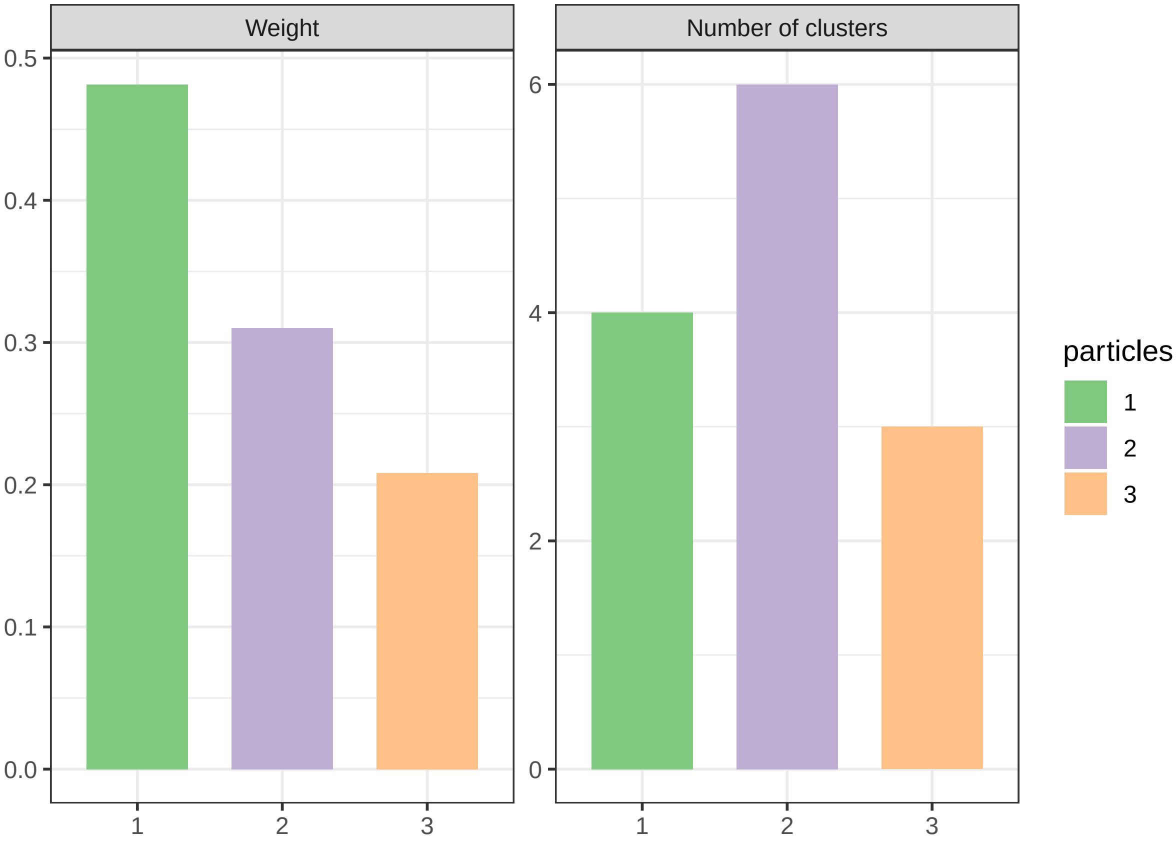
```
## Completed  1 / 5
## Completed  2 / 5
## Completed  3 / 5
## Completed  4 / 5
## Completed  5 / 5
```

```
plot(out_elbow$wass_vec, type = "b", ylab = "Wass distance", xlab = "Number of particles")
```

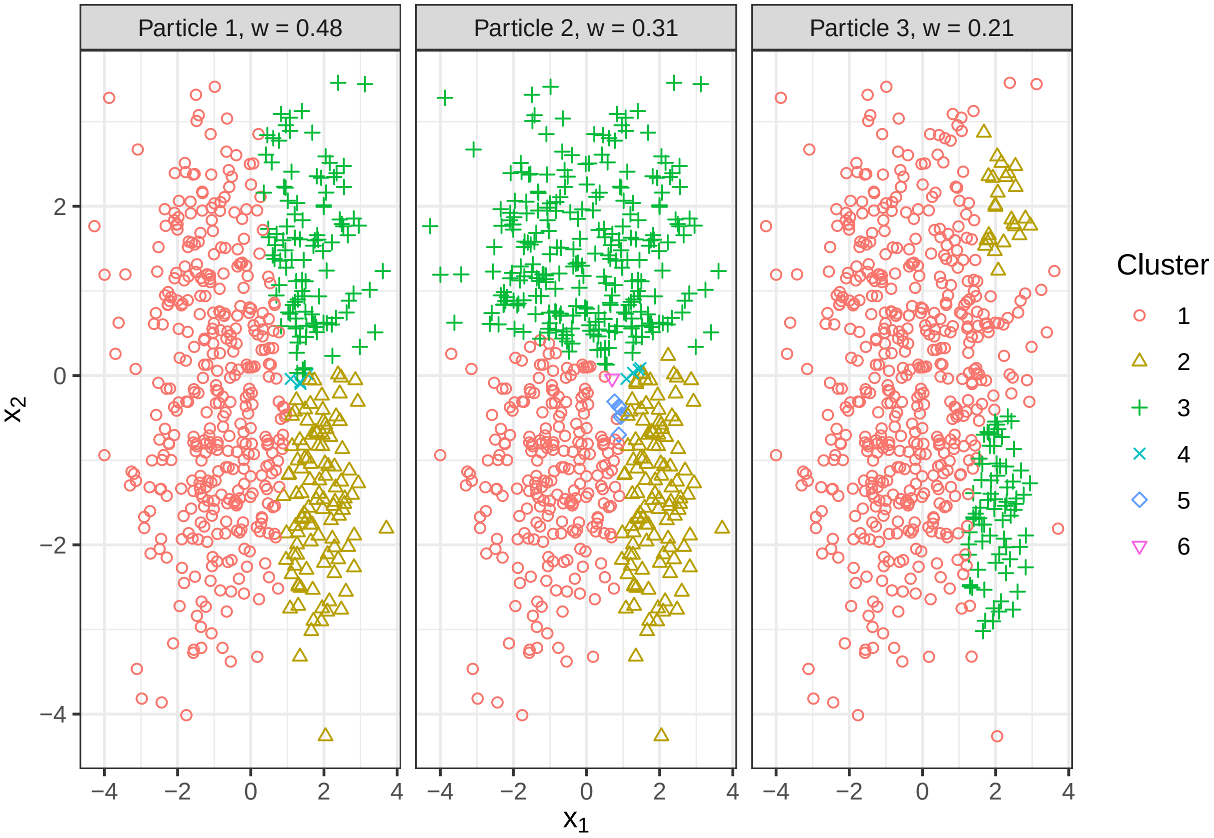


We choose "L=3" as the optimal number of clusters.

```
L = 3
output_WASABI <- out_elbow$output_list[[L]]
# output_WASABI_mb = WASABI_multistart(cls.draw, psm,
#                                      multi.start = 25, ncores = 4,
#                                      method.init ="++", add_topvi = FALSE,
#                                      method="salso", L=L,
#                                      mini.batch = 500,
#                                      max.iter= 10, extra.iter = 5,
#                                      suppress.comment=TRUE,
#                                      swap_countone = TRUE,
#                                      seed = 54321, loss = "Binder",
#                                      a = 1.8,
#                                      maxNClusters = 10)
#
# if(output_WASABI_mb$wass.dist < output_WASABI$wass.dist){
#   output_WASABI <- output_WASABI_mb
# }

ggsummary(output_WASABI)
```



```
ggscatter_grid2d(output_WASABI, Y)
```



a = 1.3

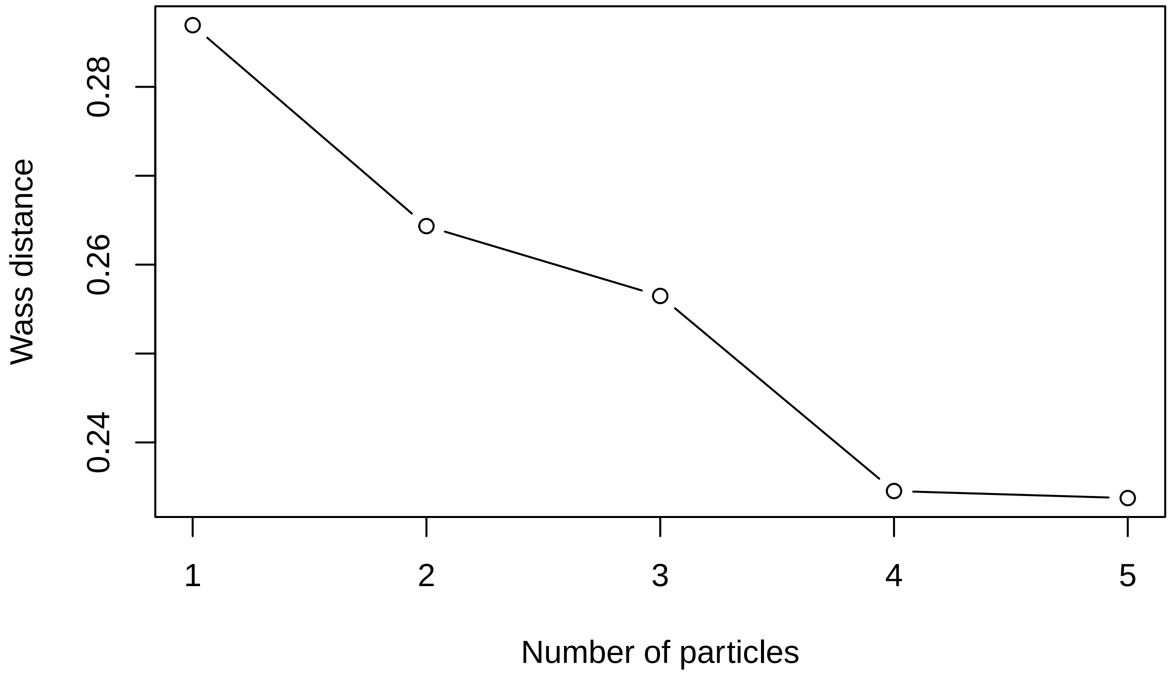
```
z_minb <- salso::salso(cls.draw, loss = binder(a = 1.3))
table(z_minb)
```

```
## z_minb
##   1   2   3
## 397 106  97
```

```
set.seed(123)
out_elbow <- elbow(cls.draw, L_max = 5, psm = psm,
                   multi.start = 1,
                   method.init = "topvi", method = "salso",
                   loss = "Binder", a = 1.3, maxNClusters = 10)
```

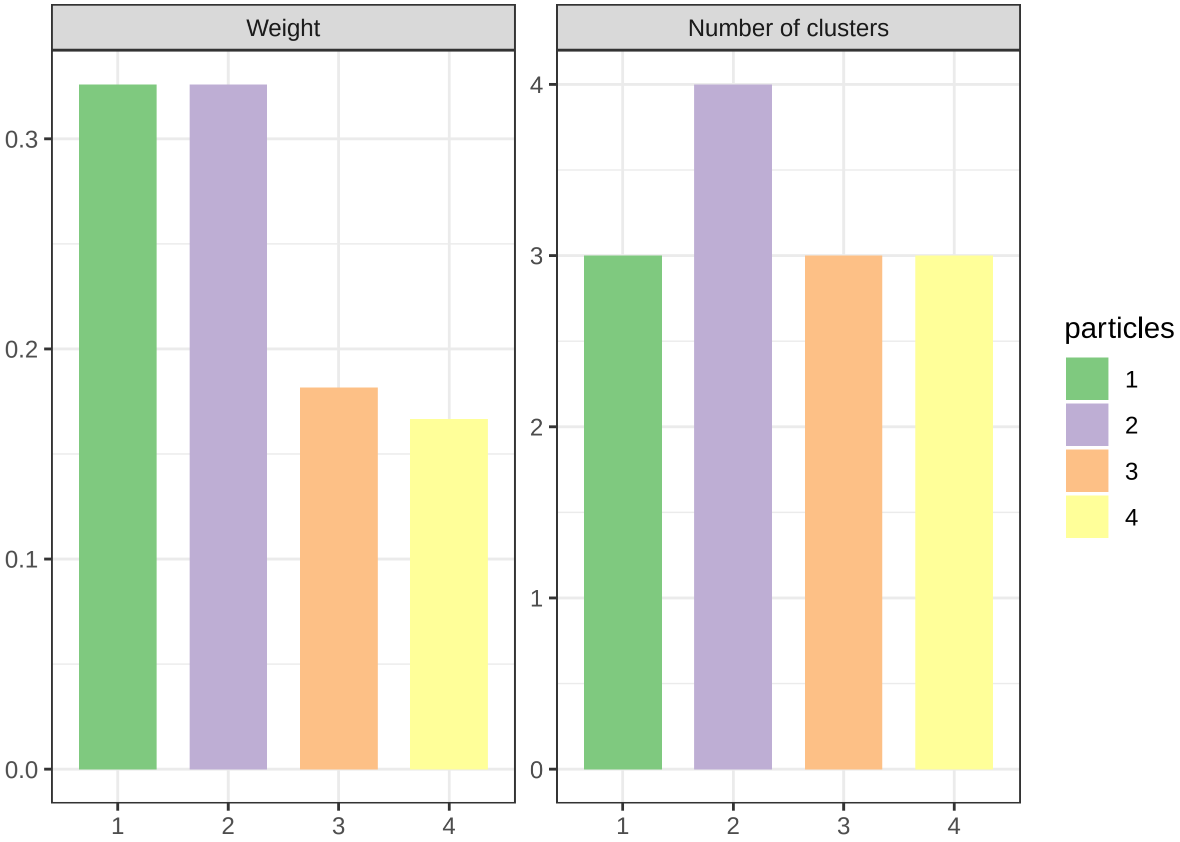
```
## Completed  1 / 5
## Completed  2 / 5
## Completed  3 / 5
## Completed  4 / 5
## Completed  5 / 5
```

```
plot(out_elbow$wass_vec, type = "b", ylab = "Wass distance", xlab = "Number of particles")
```

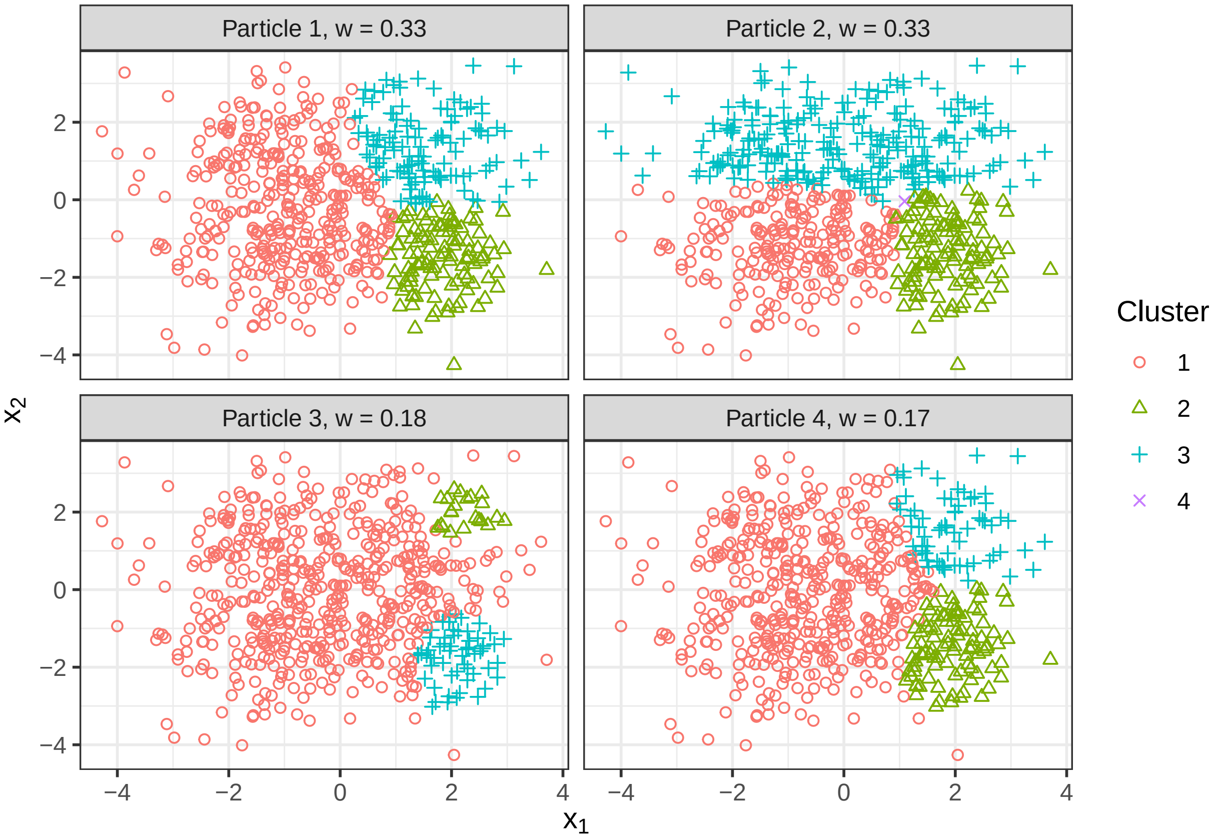


We choose “L=2” as the optimal number of clusters.

```
L = 4
output_WASABI <- out_elbow$output_list[[L]]
# output_WASABI_mb = WASABI_multistart(cls.draw, psm,
#                                      multi.start = 25, ncores = 4,
#                                      method.init ="++", add_topvi = FALSE,
#                                      method="salso", L=L,
#                                      mini.batch = 500,
#                                      max.iter= 10, extra.iter = 5,
#                                      suppress.comment=TRUE,
#                                      swap_countone = TRUE,
#                                      seed = 54321, loss = "Binder", a = 1.3,
#                                      maxNClusters = 10)
#
# if(output_WASABI_mb$wass.dist < output_WASABI$wass.dist){
#   output_WASABI <- output_WASABI_mb
# }

ggsummary(output_WASABI)
```



```
ggscatter_grid2d(output_WASABI, Y)
```



a = 1.5

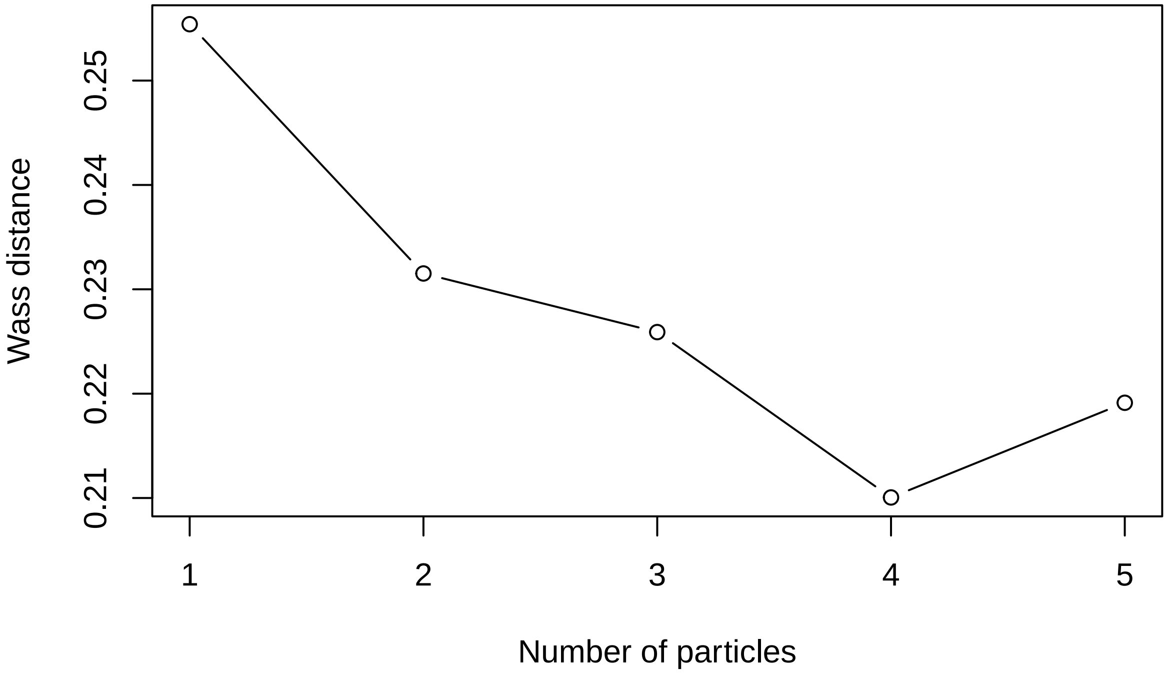
```
z_minb <- salso::salso(cls.draw, loss = binder(a = 1.5))
table(z_minb)
```

```
## z_minb
##   1   2
## 506  94
```

```
set.seed(123)
out_elbow <- elbow(cls.draw, L_max = 5, psm = psm,
                   multi.start = 1,
                   method.init = "topvi", method = "salso",
                   loss = "Binder", a = 1.5, maxNClusters = 10)
```

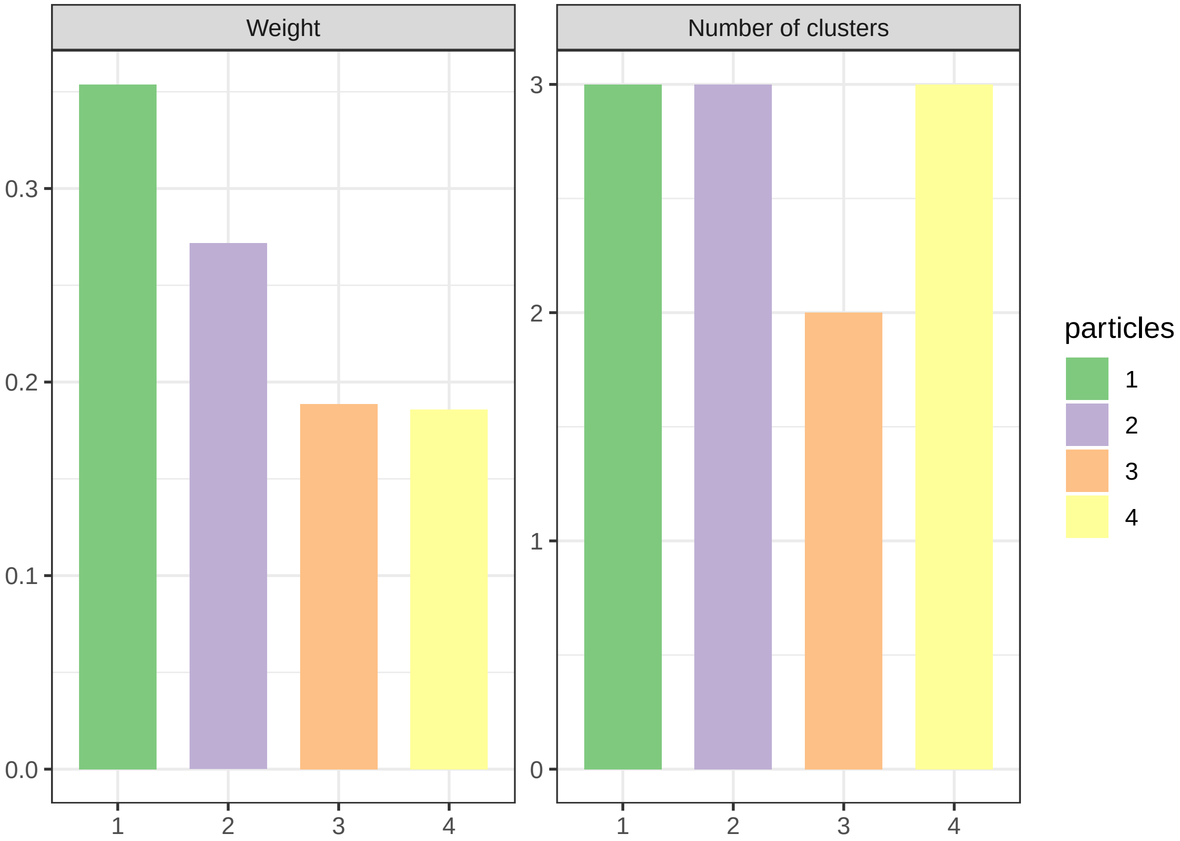
```
## Completed  1 / 5
## Completed  2 / 5
## Completed  3 / 5
## Completed  4 / 5
## Completed  5 / 5
```

```
plot(out_elbow$wass_vec, type = "b", ylab = "Wass distance", xlab = "Number of particles")
```

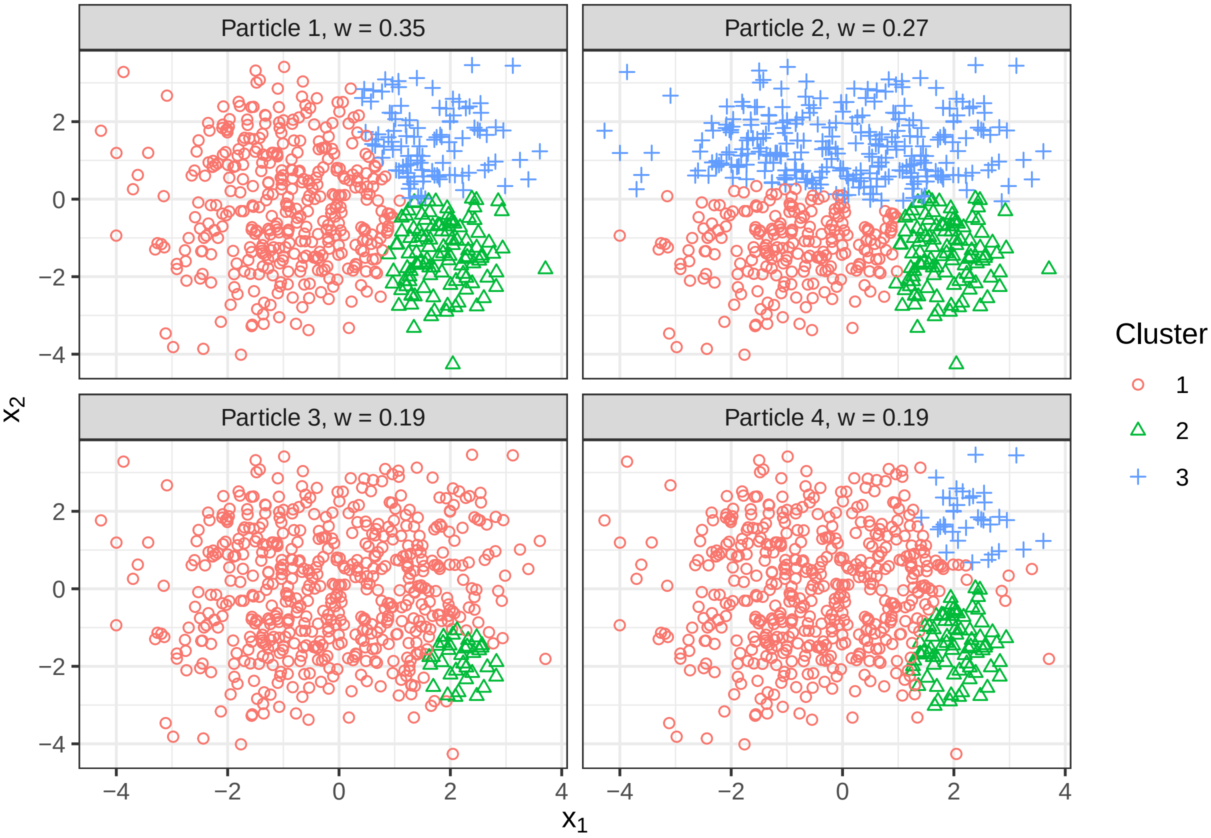


We choose “L=2” as the optimal number of clusters.

```
L = 4
output_WASABI <- out_elbow$output_list[[L]]
# output_WASABI_mb = WASABI_multistart(cls.draw, psm,
#                                      multi.start = 25, ncores = 4,
#                                      method.init ="++", add_topvi = FALSE,
#                                      method="salso", L=L,
#                                      mini.batch = 500,
#                                      max.iter= 10, extra.iter = 5,
#                                      suppress.comment=TRUE,
#                                      swap_countone = TRUE,
#                                      seed = 54321, loss = "Binder", a = 1.2,
#                                      maxNClusters = 10)
#
# if(output_WASABI_mb$wass.dist < output_WASABI$wass.dist){
#   output_WASABI <- output_WASABI_mb
# }

ggsummary(output_WASABI)
```



```
ggscatter_grid2d(output_WASABI, Y)
```


Particle 1, w = 0.35
Particle 2, w = 0.27
Particle 3, w = 0.19
Particle 4, w = 0.19
x2
x1
2
0
−2
−4
−4
−2
0
2
4
Cluster
○ 1
△ 2
+ 3