Binder

Guanyu

2025-08-10

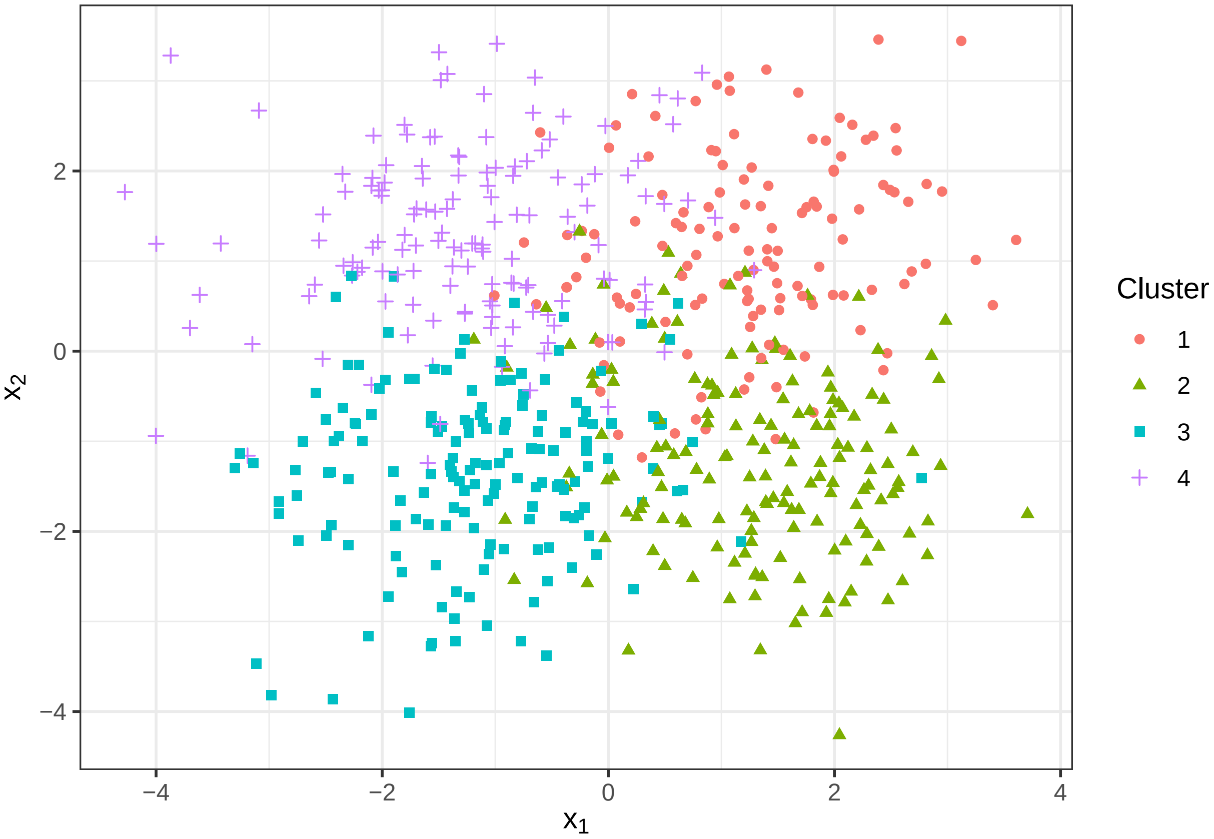
Two-dimensional data

```
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
```



Run MCMC

```
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 0.525079 sec
## Completed:   3000/15000 - in 1.0177 sec
## Completed:   4500/15000 - in 1.49858 sec
## Completed:   6000/15000 - in 2.12987 sec
## Completed:   7500/15000 - in 2.74801 sec
## Completed:   9000/15000 - in 3.43799 sec
## Completed:   10500/15000 - in 4.14226 sec
## Completed:   12000/15000 - in 4.80781 sec
## Completed:   13500/15000 - in 5.44719 sec
## Completed:   15000/15000 - in 6.11111 sec
##
## Estimation done in 6.11116 seconds
```

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

Parameter selction

Inspired by salso paper's experiment, we can roughly devide the range of a with respect to number of clusters.

For $a \in [1.065, 1.125)$ there are 6 clusters produced.

```
z_minb1 <- salso::salso(cls.draw, loss = binder(a = 1.065))
table(z_minb1)
```

```
## z_minb1
##   1   2   3   4   5   6
## 345  99  14 110   6  26
```

For $a \in [1.125, 1.168)$ there are 5 clusters produced.

```
z_minb2 <- salso::salso(cls.draw, loss = binder(a = 1.13), maxNClusters = 10)
table(z_minb2)
```

```
## z_minb2
##   1   2   3   4   5
## 357 100  16 107  20
```

For $a \in [1.168, 1.213)$ there are 4 clusters produced.

```
z_minb3 <- salso::salso(cls.draw, loss = binder(a = 1.17), maxNClusters = 10)
table(z_minb3)
```

```
## z_minb3
##   1   2   3   4
## 384 105 104   7
```

For $a \in [1.213, 1.47)$ there are 3 clusters produced.

```
z_minb4 <- salso::salso(cls.draw, loss = binder(a = 1.3), maxNClusters = 10)
table(z_minb4)
```

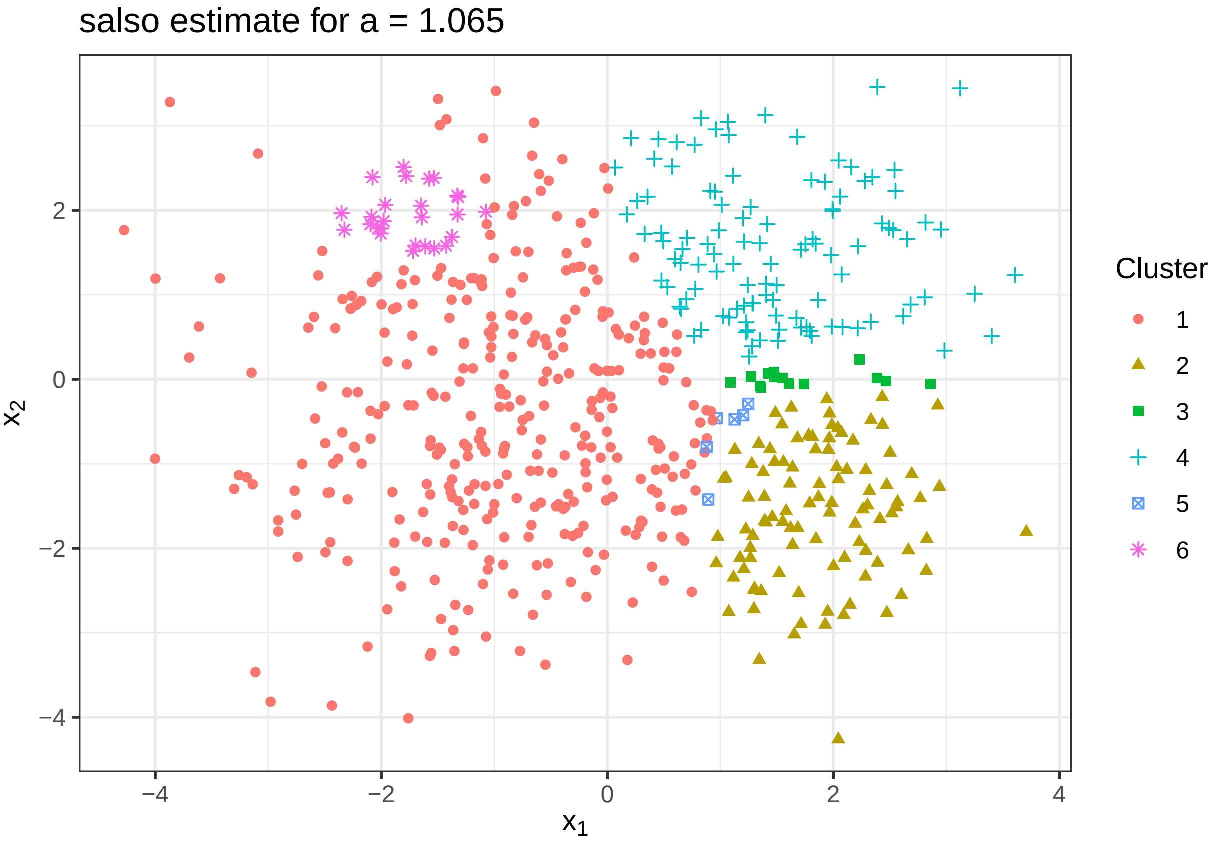
```
## z_minb4
##   1   2   3
## 397 106  97
```

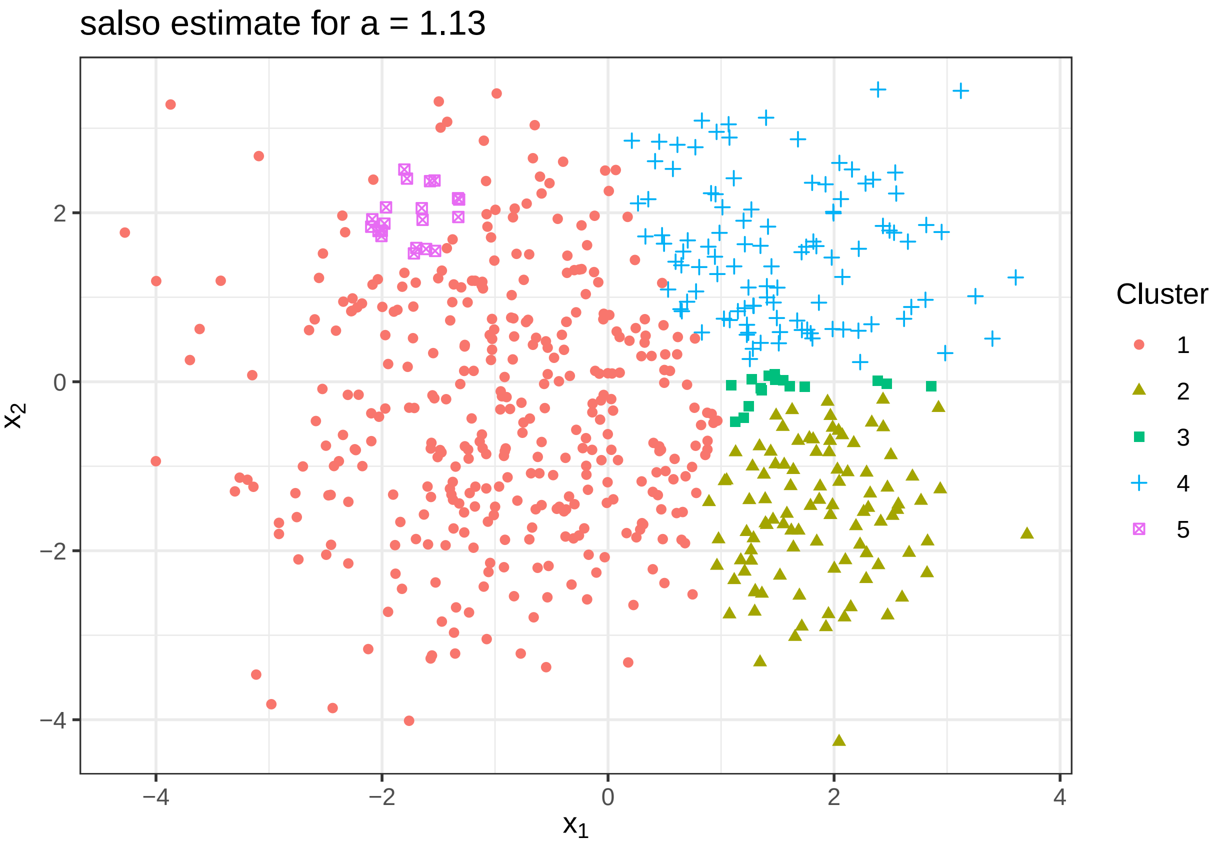
For $a \in [1.47, 1.665)$ there are 2 clusters produced. And for a between 1.665 and 2, only one cluster is produced.

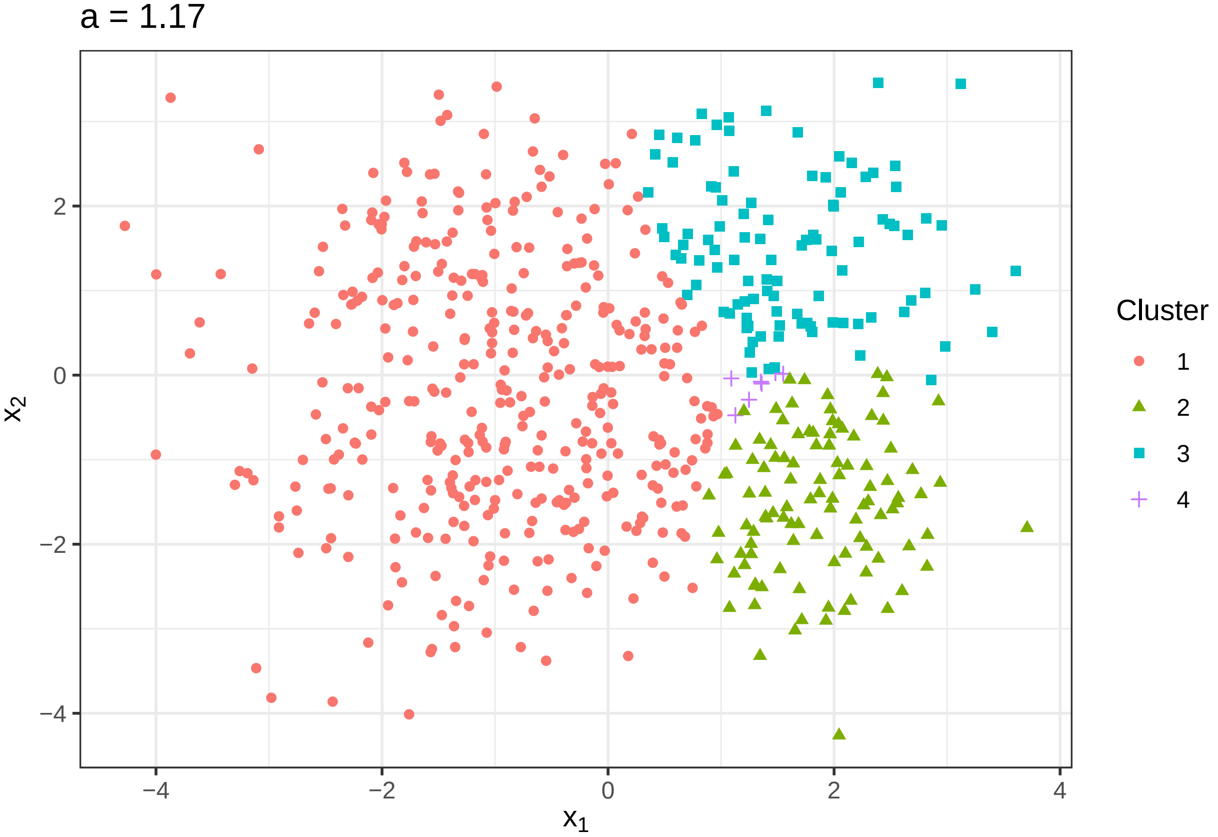
```
z_minb5 <- salso::salso(cls.draw, loss = binder(a = 1.5), maxNClusters = 10)
table(z_minb5)
```

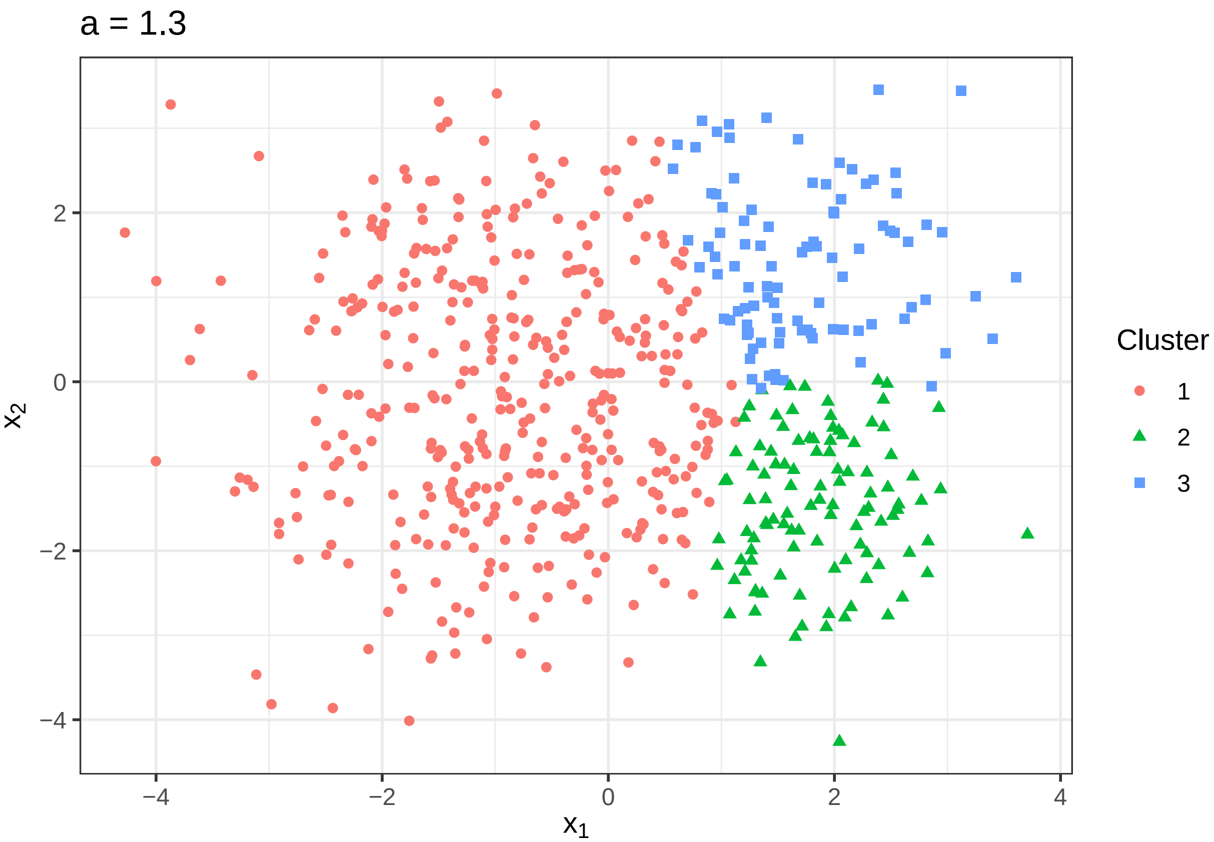
```
## z_minb5
##   1   2
## 506  94
```

We can put all the plots together to see how the number of clusters change.

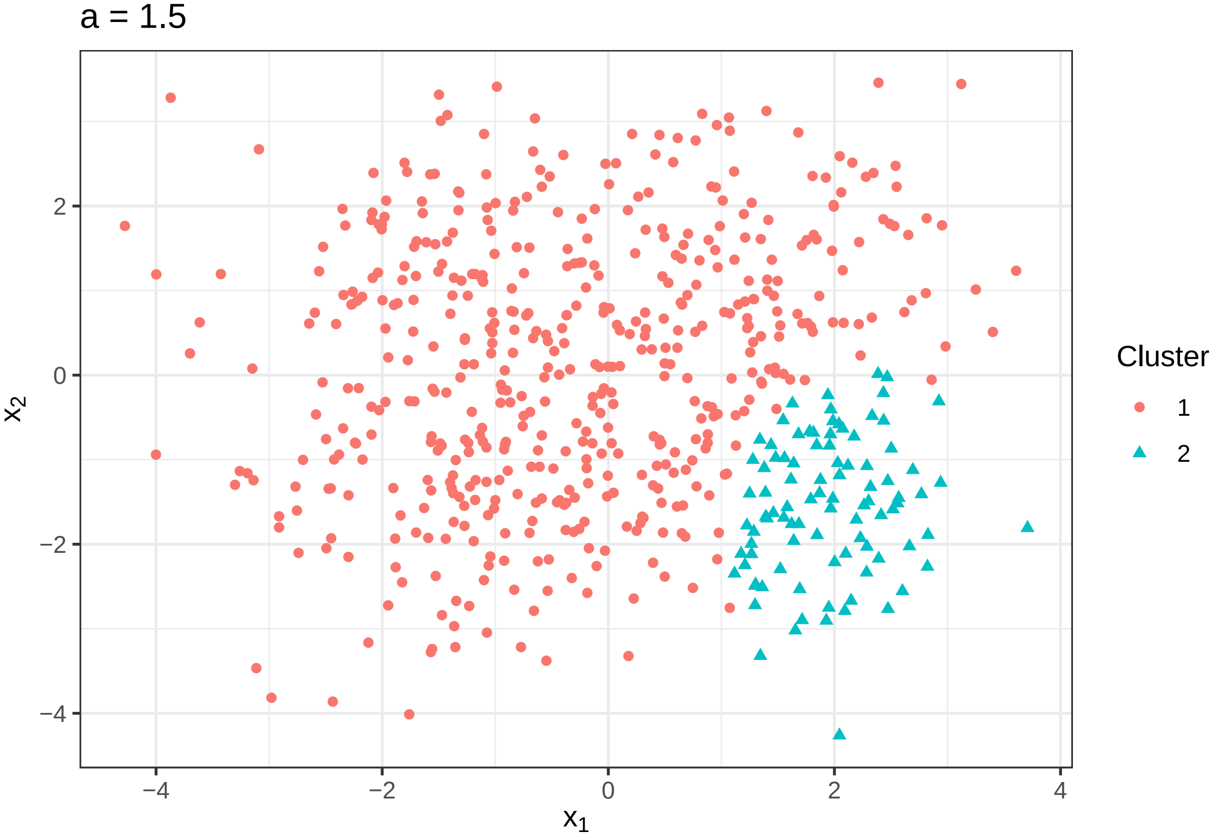



salso estimate for a = 1.13
x2
2
0
−2
−4
−4
−2
0
2
4
x1
Cluster
1
2
3
4
5


a = 1.17
x2
2
0
−2
−4
−4
−2
0
2
4
x1
Cluster
1
2
3
4



a = 1.3



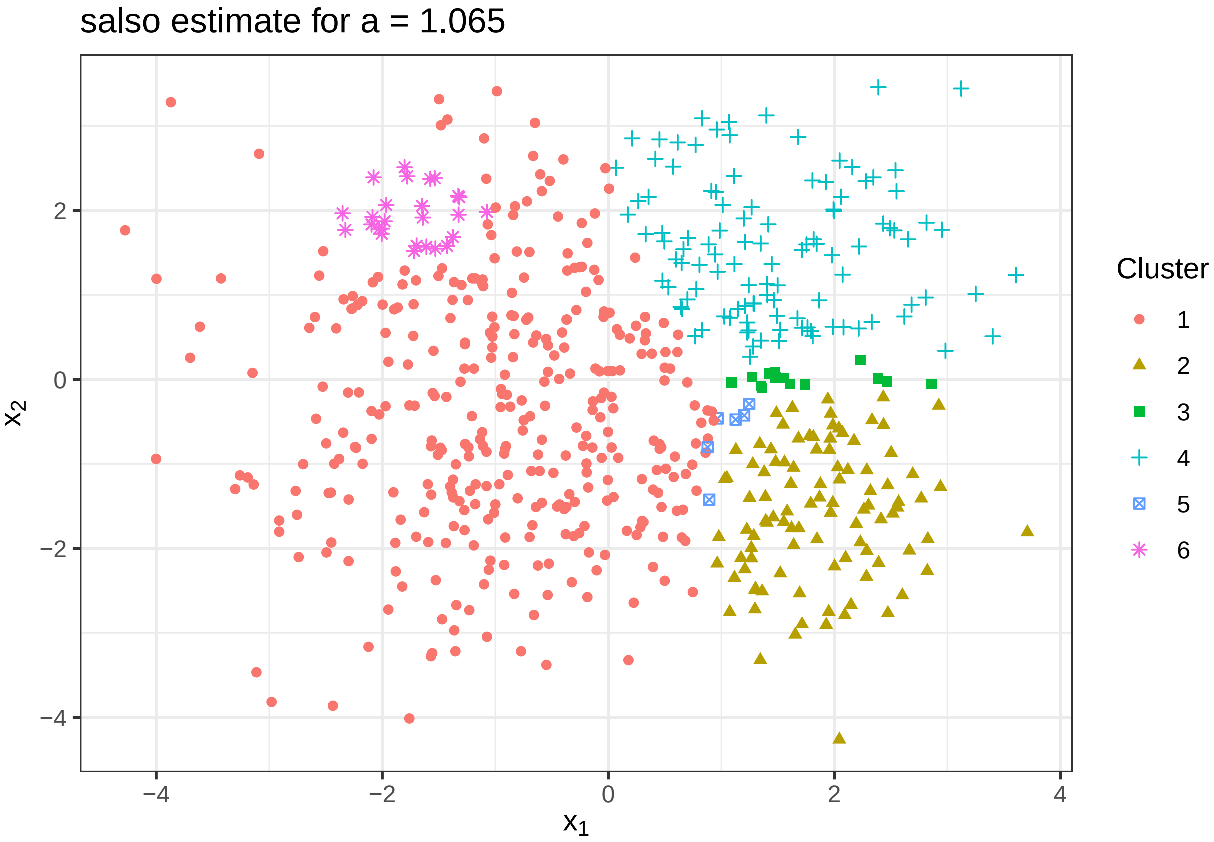
a = 1.5

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

a = 1.065

```
table(z_minb1)
```

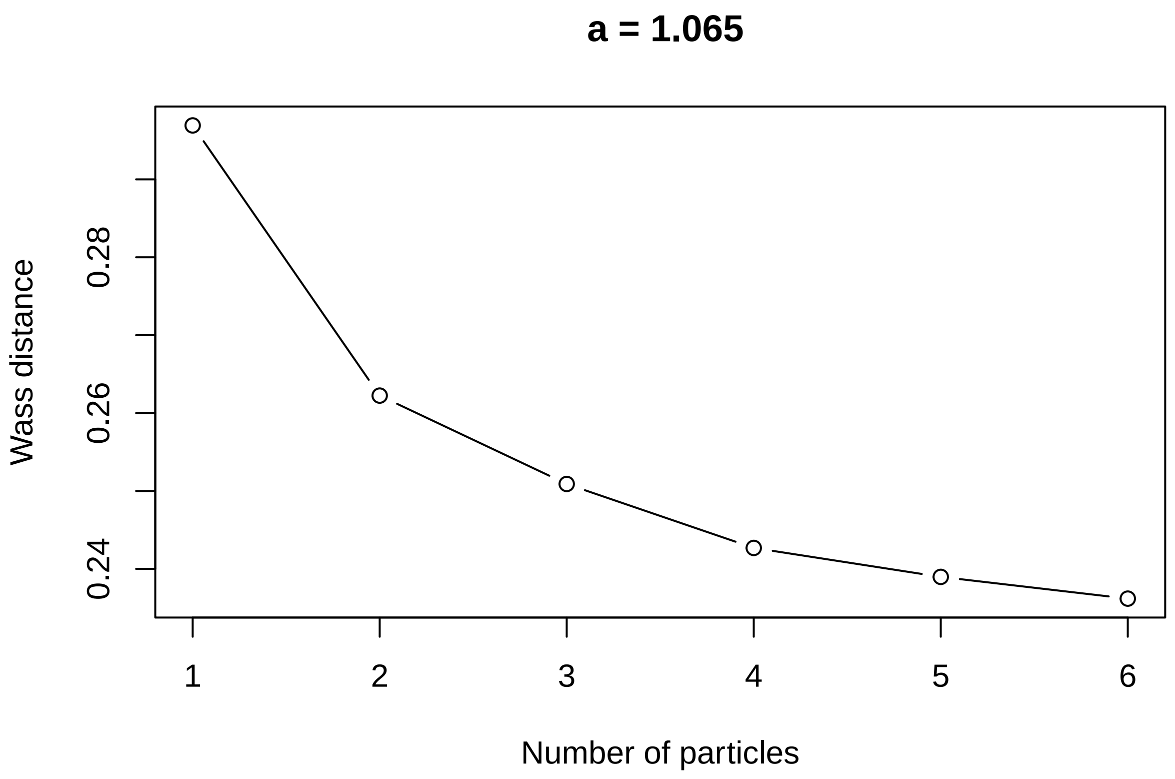
```
## z_minb1
##   1   2   3   4   5   6
## 345  99  14 110   6  26
```



```
set.seed(123)
out_elbow <- elbow(cls.draw, L_max = 6, psm = psm,
                   multi.start = 6, method.init = "++",
                   method = "salso", mini.batch = 500, ncores = 6,
                   loss = "Binder", a = 1.065, maxNClusters = 10)
```

```
## Completed  1 / 6
## Completed  2 / 6
## Completed  3 / 6
## Completed  4 / 6
## Completed  5 / 6
## Completed  6 / 6
```

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

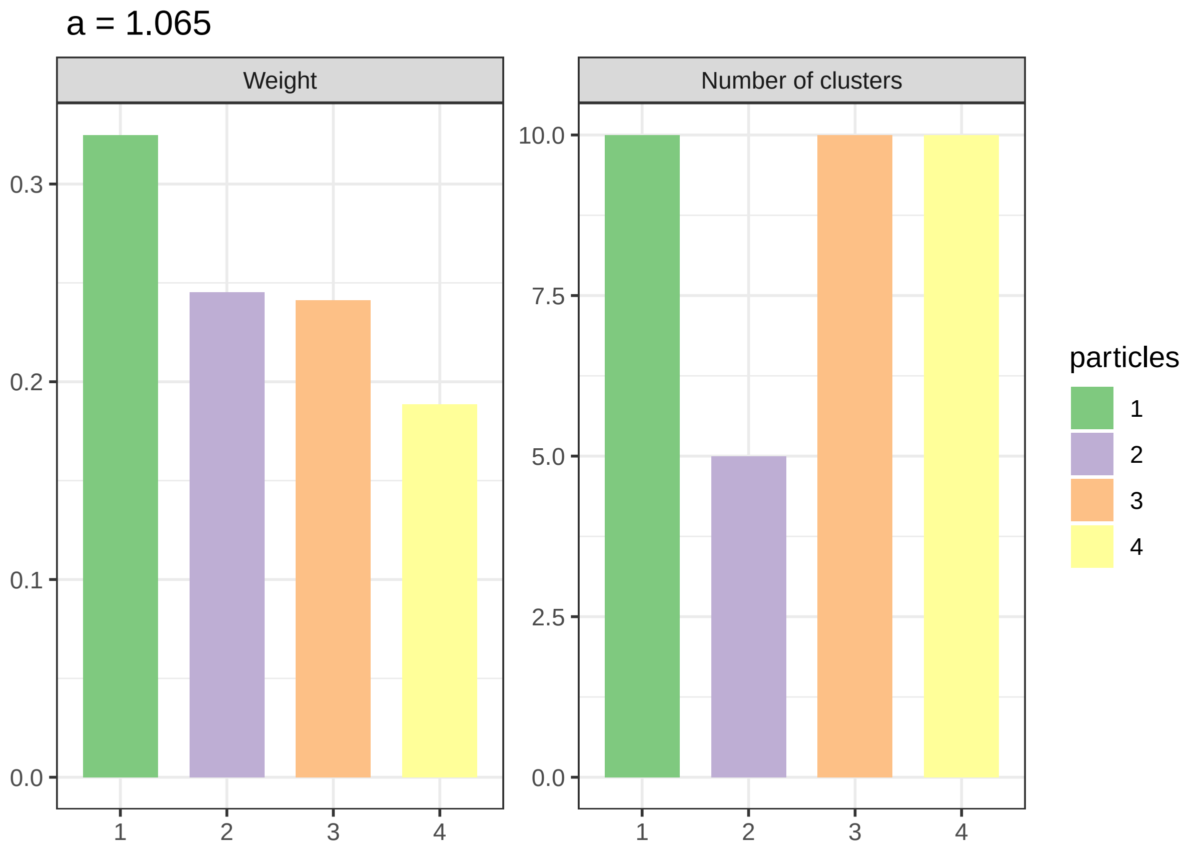


We choose “L=4” as the optimal number of clusters.

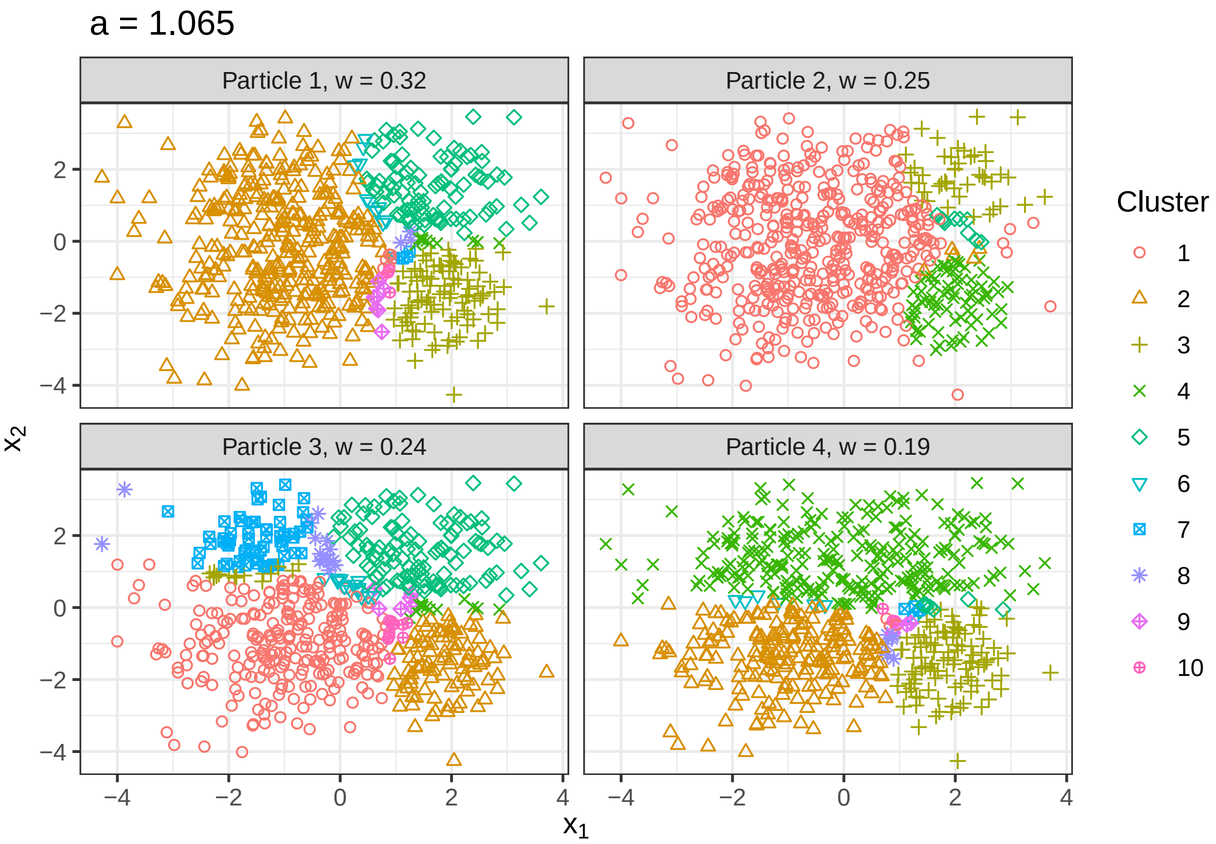
```
L = 4
output_WASABI <- out_elbow$output_list[[L]]
output_WASABI_mb = WASABI_multistart(cls.draw, psm,
                                    multi.start = 25, ncores = 6,
                                    method.init ="++", add_topvi = FALSE,
                                    method="salso", L=L,
                                    mini.batch = 500,
                                    max.iter= 10, extra.iter = 5,
                                    suppress.comment=TRUE,
                                    swap_countone = TRUE,
                                    seed = 54321, loss = "Binder",
                                    a = 1.065,
                                    maxNClusters = 10)

if(output_WASABI_mb$wass.dist < output_WASABI$wass.dist){
  output_WASABI <- output_WASABI_mb
}

ggsummary(output_WASABI, title = " a = 1.065 ")
```



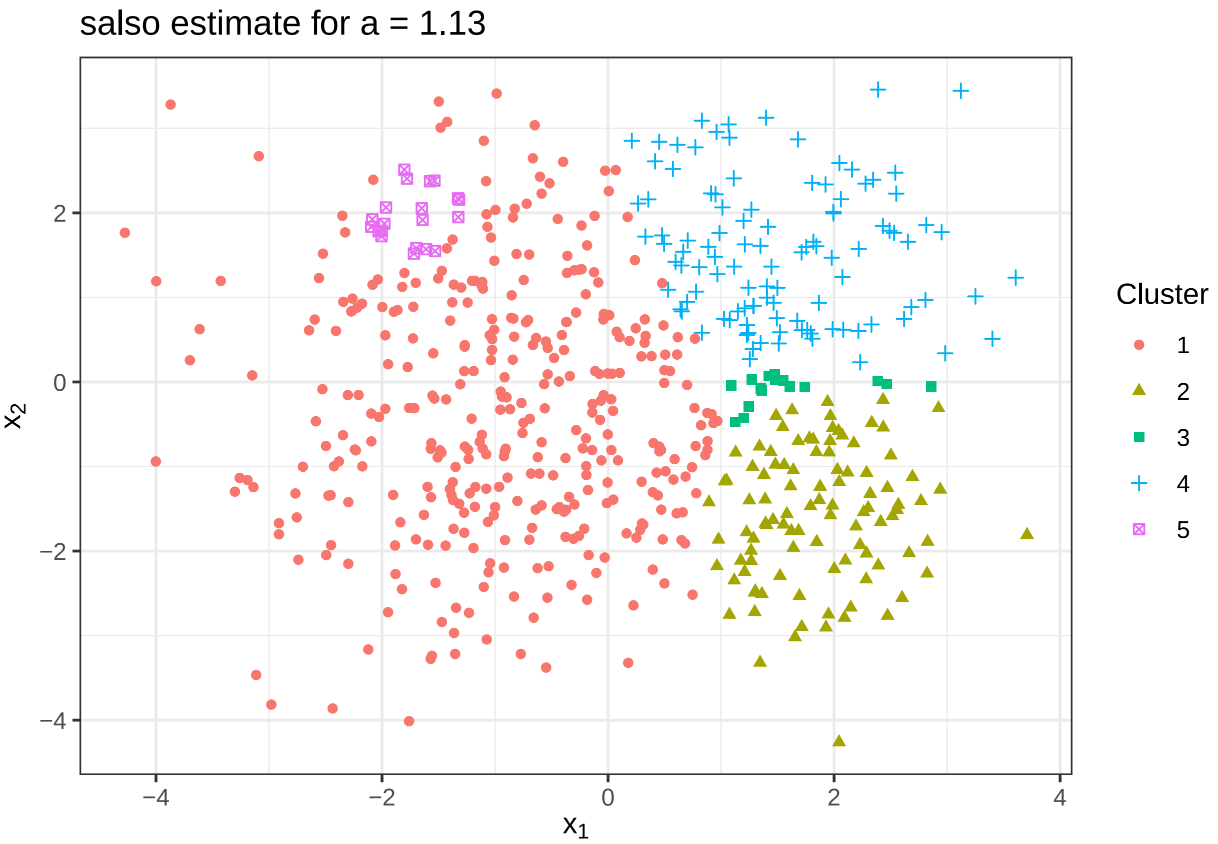
```
ggscatter_grid2d(output_WASABI, Y, title = " a = 1.065 ")
```



a = 1.13

```
table(z_minb2)
```

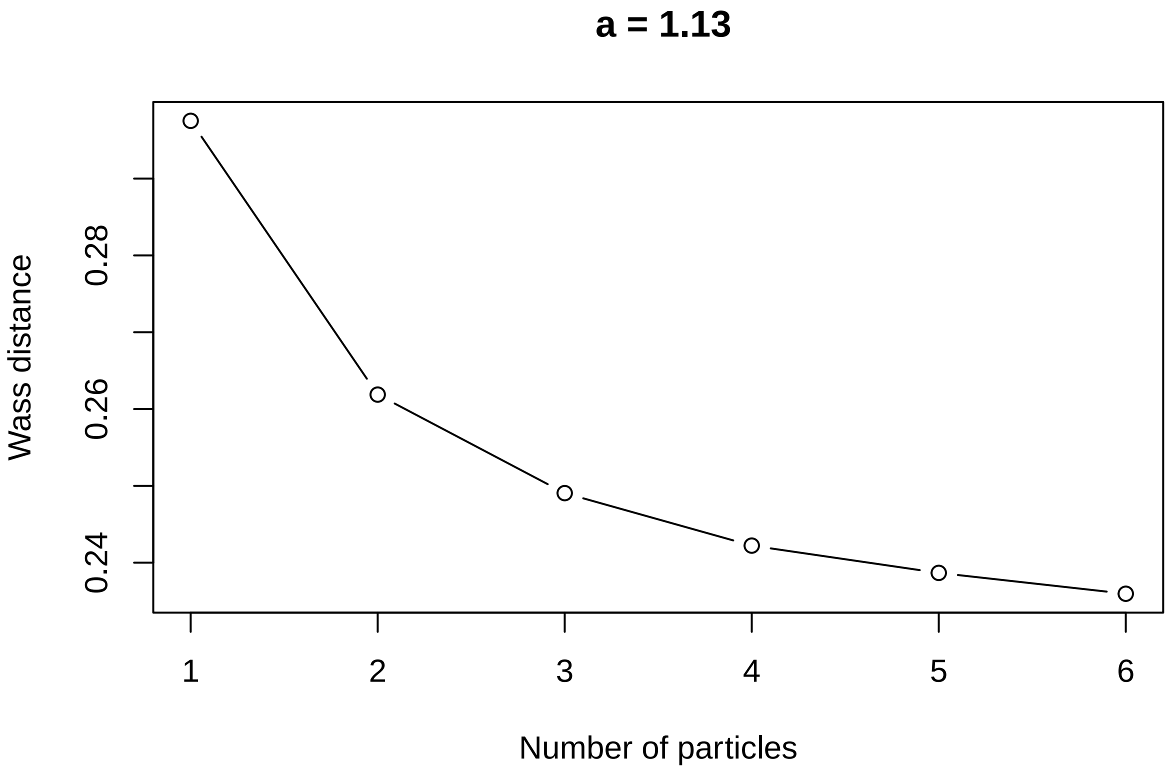
```
## z_minb2
##   1   2   3   4   5
## 357 100  16 107  20
```



```r
set.seed(123)
out_elbow <- elbow(cls.draw, L_max = 6, psm = psm,
                   multi.start = 6, method.init = "++",
                   method = "salso", mini.batch = 500, ncores = 6,
                   loss = "Binder", a = 1.13, maxNClusters = 10)
```

```
## Completed  1 / 6
## Completed  2 / 6
## Completed  3 / 6
## Completed  4 / 6
## Completed  5 / 6
## Completed  6 / 6
```

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

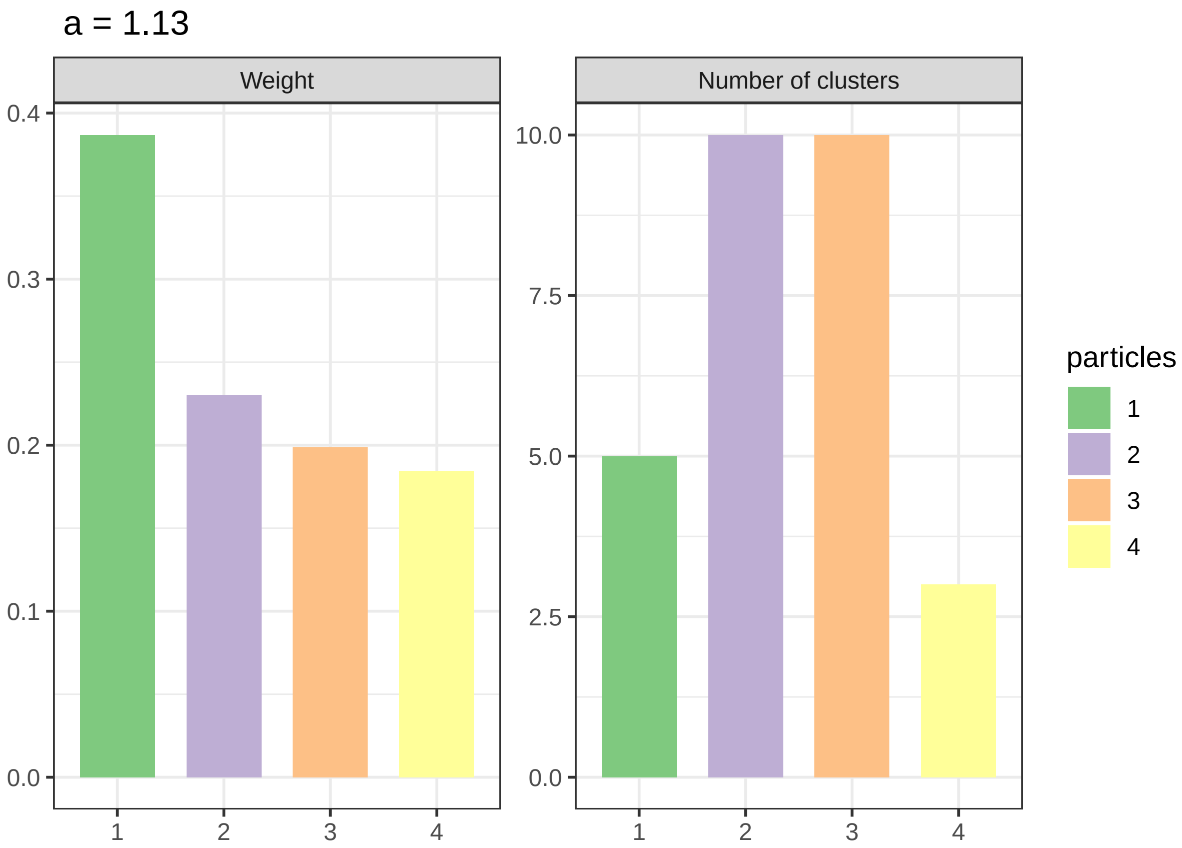


We choose “L=4” as the optimal number of clusters.

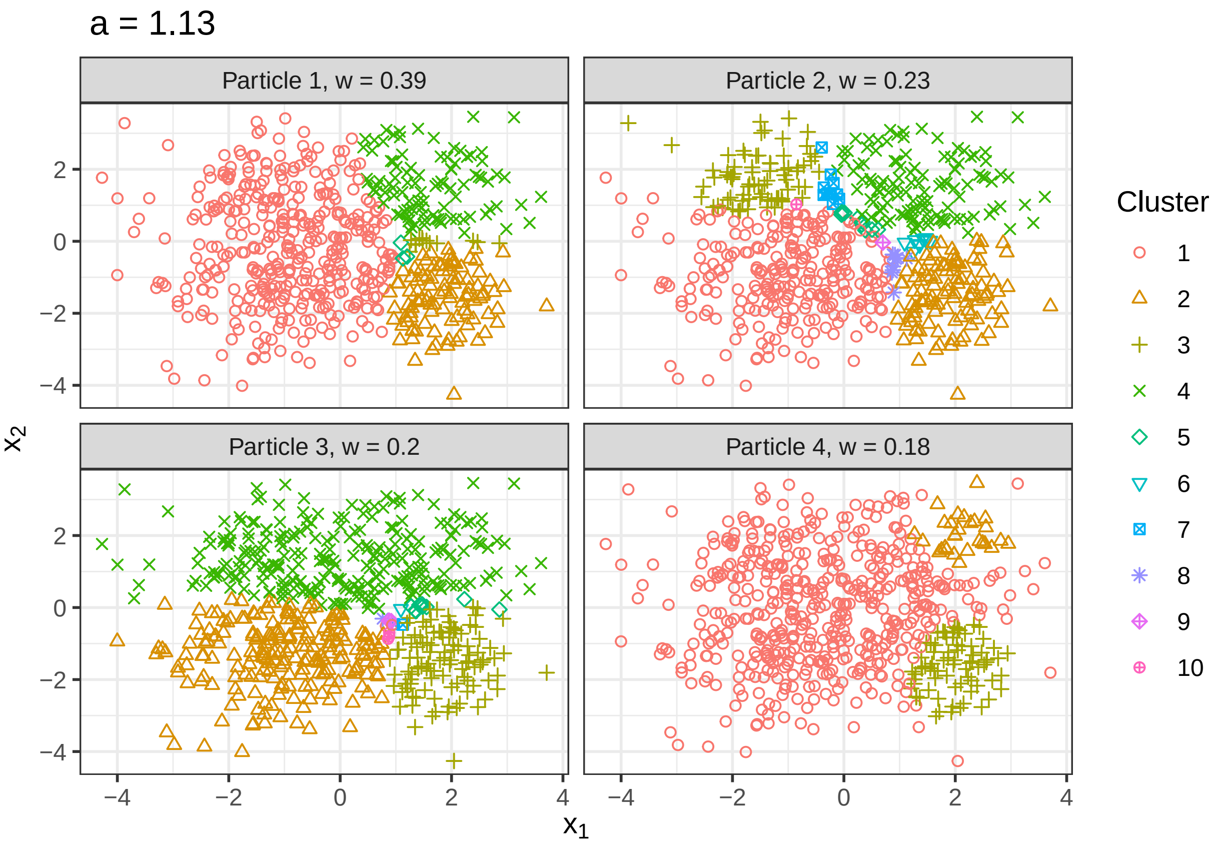
```
L = 4
output_WASABI <- out_elbow$output_list[[L]]
output_WASABI_mb = WASABI_multistart(cls.draw, psm,
                                    multi.start = 25, ncores = 6,
                                    method.init ="++", add_topvi = FALSE,
                                    method="salso", L=L,
                                    mini.batch = 500,
                                    max.iter= 10, extra.iter = 5,
                                    suppress.comment=TRUE,
                                    swap_countone = TRUE,
                                    seed = 54321, loss = "Binder",
                                    a = 1.13,
                                    maxNClusters = 10)

if(output_WASABI_mb$wass.dist < output_WASABI$wass.dist){
  output_WASABI <- output_WASABI_mb
}

ggsummary(output_WASABI, title = " a = 1.13 ")
```



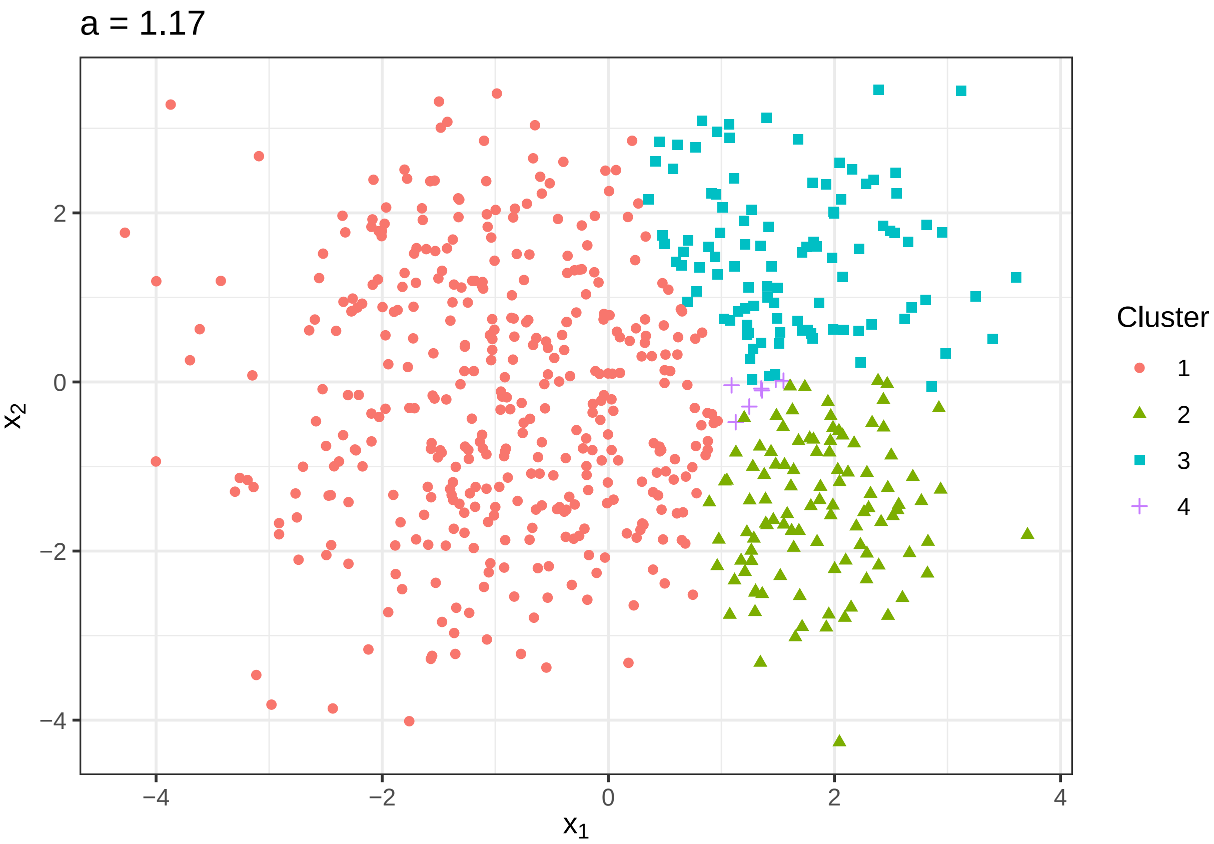
```
ggscatter_grid2d(output_WASABI, Y, title = " a = 1.13 ")
```



a = 1.17

```
table(z_minb3)
```

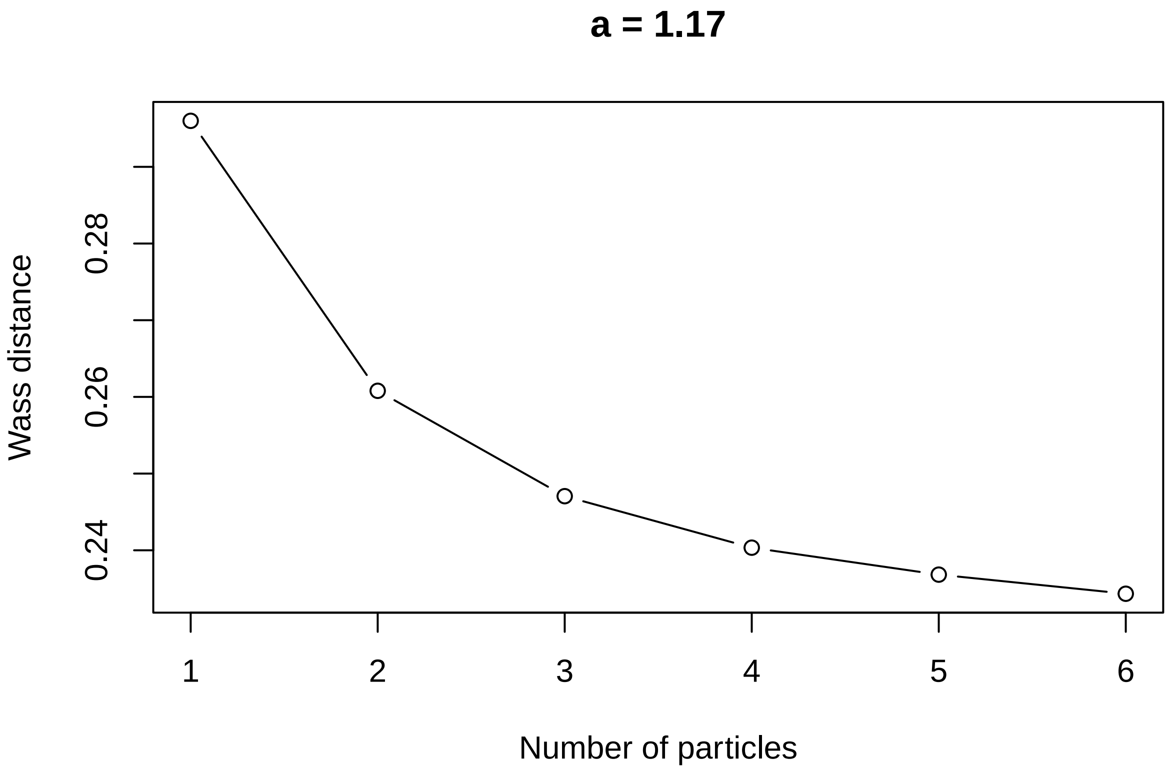
```
## z_minb3
##   1   2   3   4
## 384 105 104   7
```



```
set.seed(123)
out_elbow <- elbow(cls.draw, L_max = 6, psm = psm,
                   multi.start = 6, method.init = "++",
                   method = "salso", mini.batch = 500, ncores = 6,
                   loss = "Binder", a = 1.17, maxNClusters = 10)
```

```
## Completed  1 / 6
## Completed  2 / 6
## Completed  3 / 6
## Completed  4 / 6
## Completed  5 / 6
## Completed  6 / 6
```

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

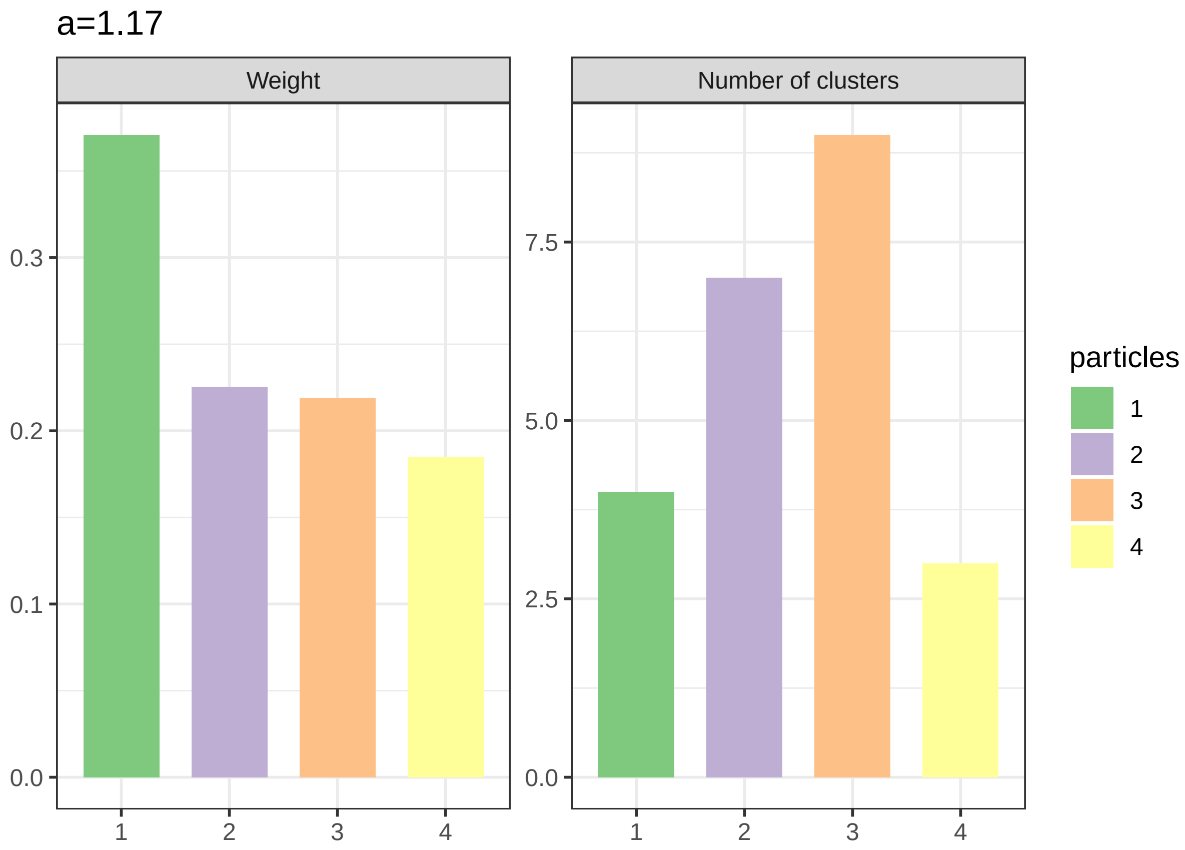


We choose “L=4” as the optimal number of clusters.

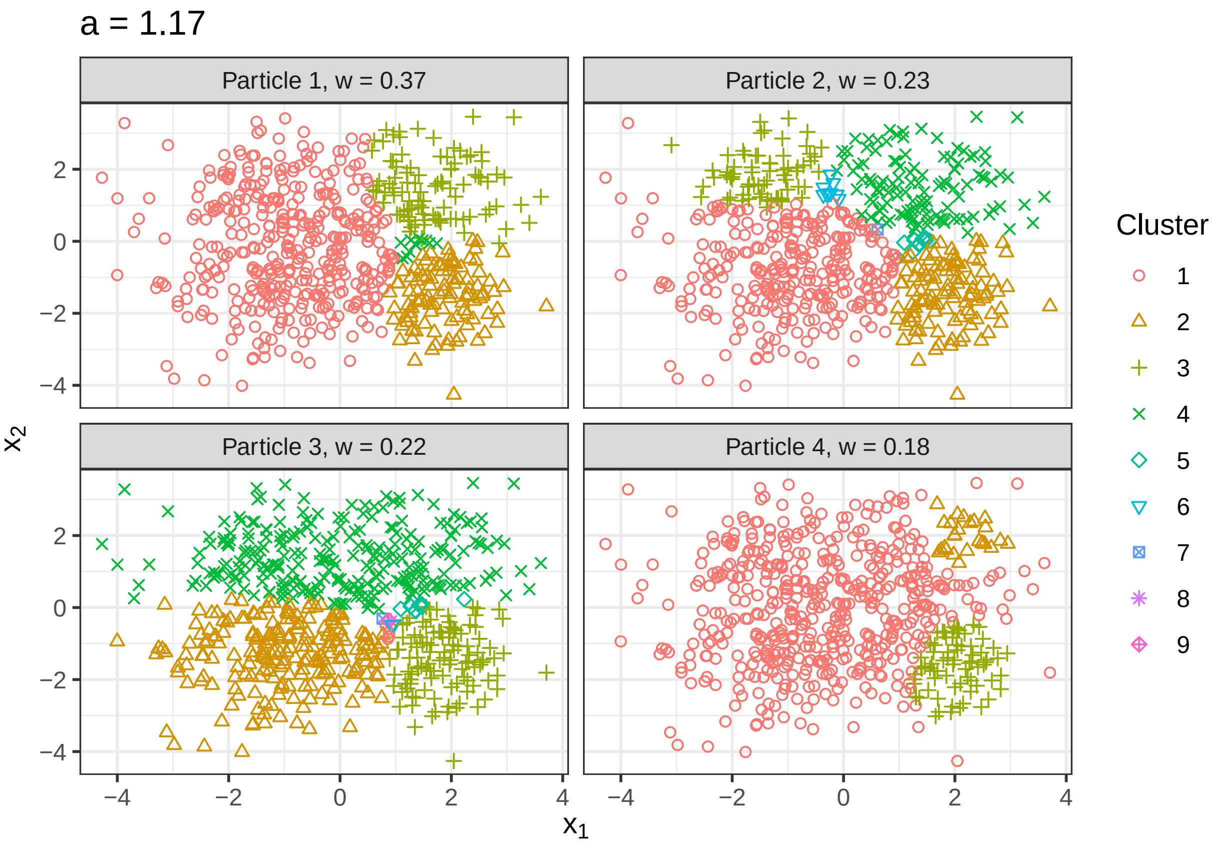
```
L = 4
output_WASABI <- out_elbow$output_list[[L]]
output_WASABI_mb = WASABI_multistart(cls.draw, psm,
                                    multi.start = 25, ncores = 6,
                                    method.init ="++", add_topvi = FALSE,
                                    method="salso", L=L,
                                    mini.batch = 500,
                                    max.iter= 10, extra.iter = 5,
                                    suppress.comment=TRUE,
                                    swap_countone = TRUE,
                                    seed = 54321, loss = "Binder", a = 1.17,
                                    maxNClusters = 10)

if(output_WASABI_mb$wass.dist < output_WASABI$wass.dist){
  output_WASABI <- output_WASABI_mb
}

ggsummary(output_WASABI, title = "a=1.17")
```



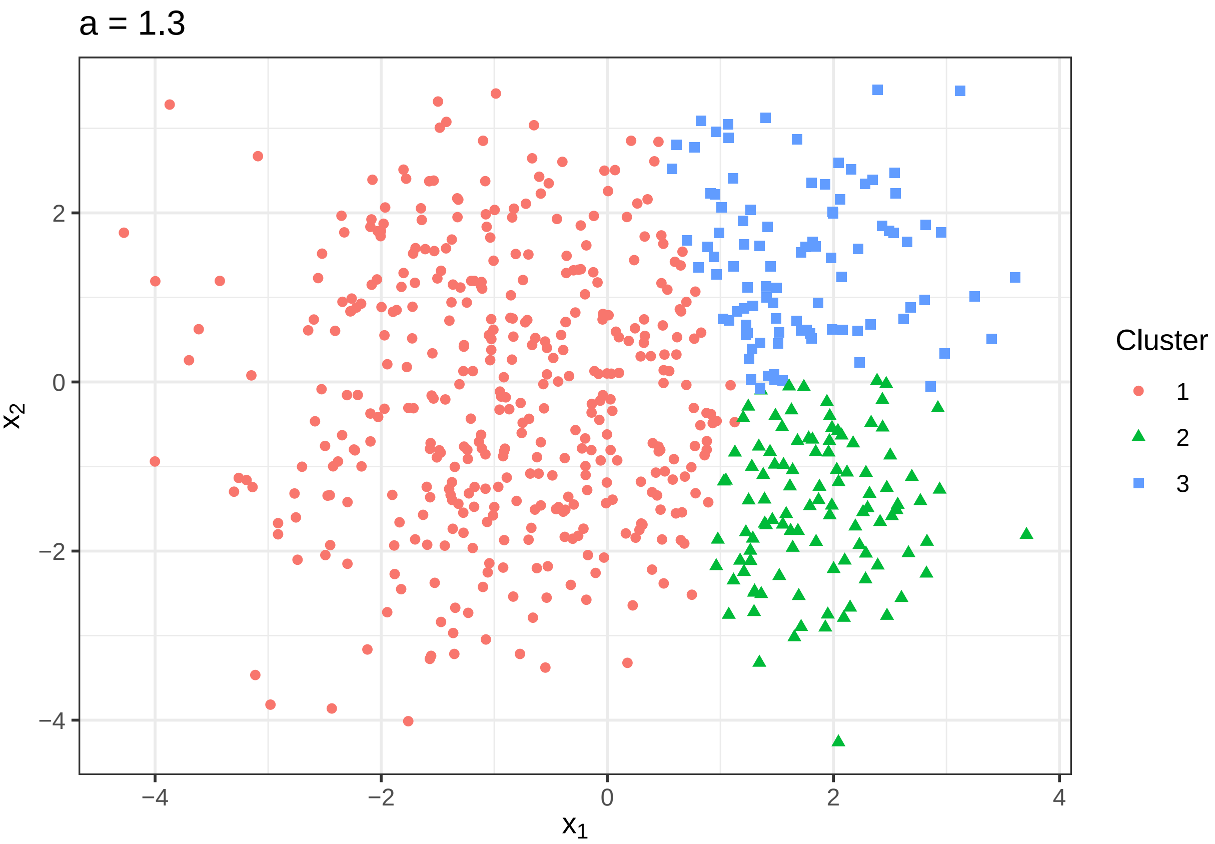
```
ggscatter_grid2d(output_WASABI, Y, title = "a = 1.17 ")
```



a = 1.3

```r
table(z_minb4)
```

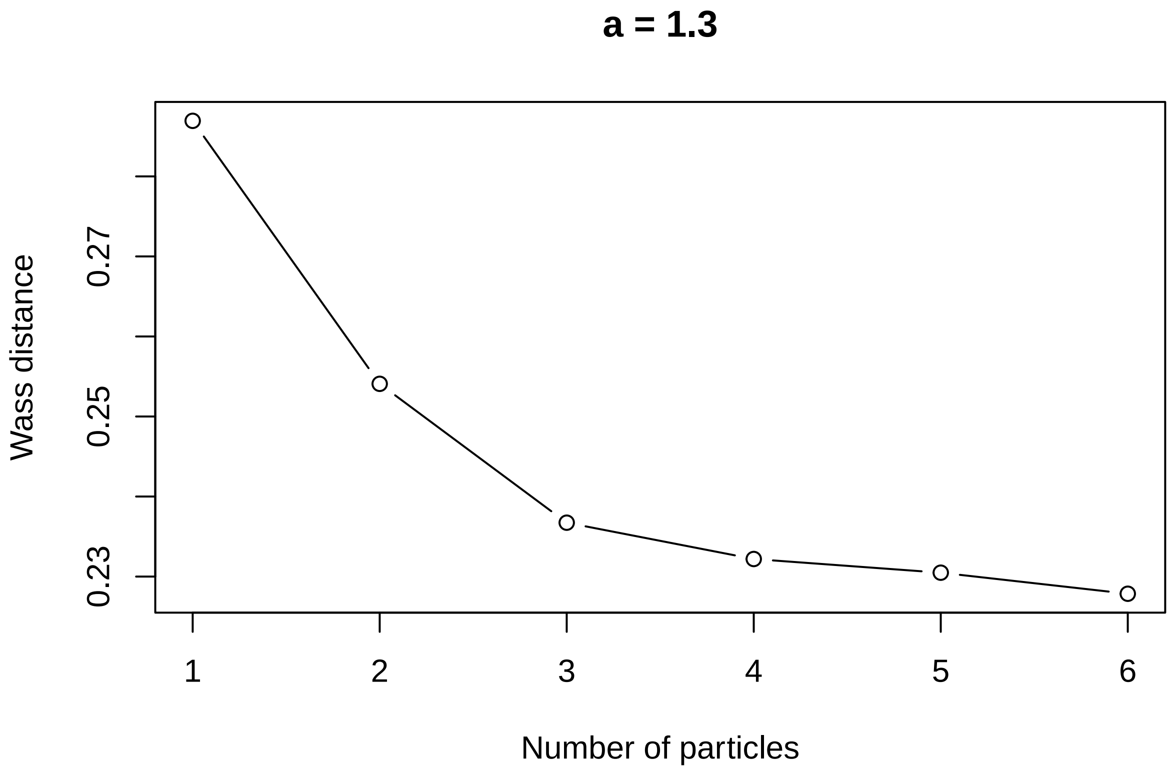
```
## z_minb4
##   1   2   3
## 397 106  97
```



```
set.seed(123)
out_elbow <- elbow(cls.draw, L_max = 6, psm = psm,
                   multi.start = 6, method.init = "++",
                   method = "salso", mini.batch = 500, ncores = 6,
                   loss = "Binder", a = 1.3, maxNClusters = 10)
```

```
## Completed  1 / 6
## Completed  2 / 6
## Completed  3 / 6
## Completed  4 / 6
## Completed  5 / 6
## Completed  6 / 6
```

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

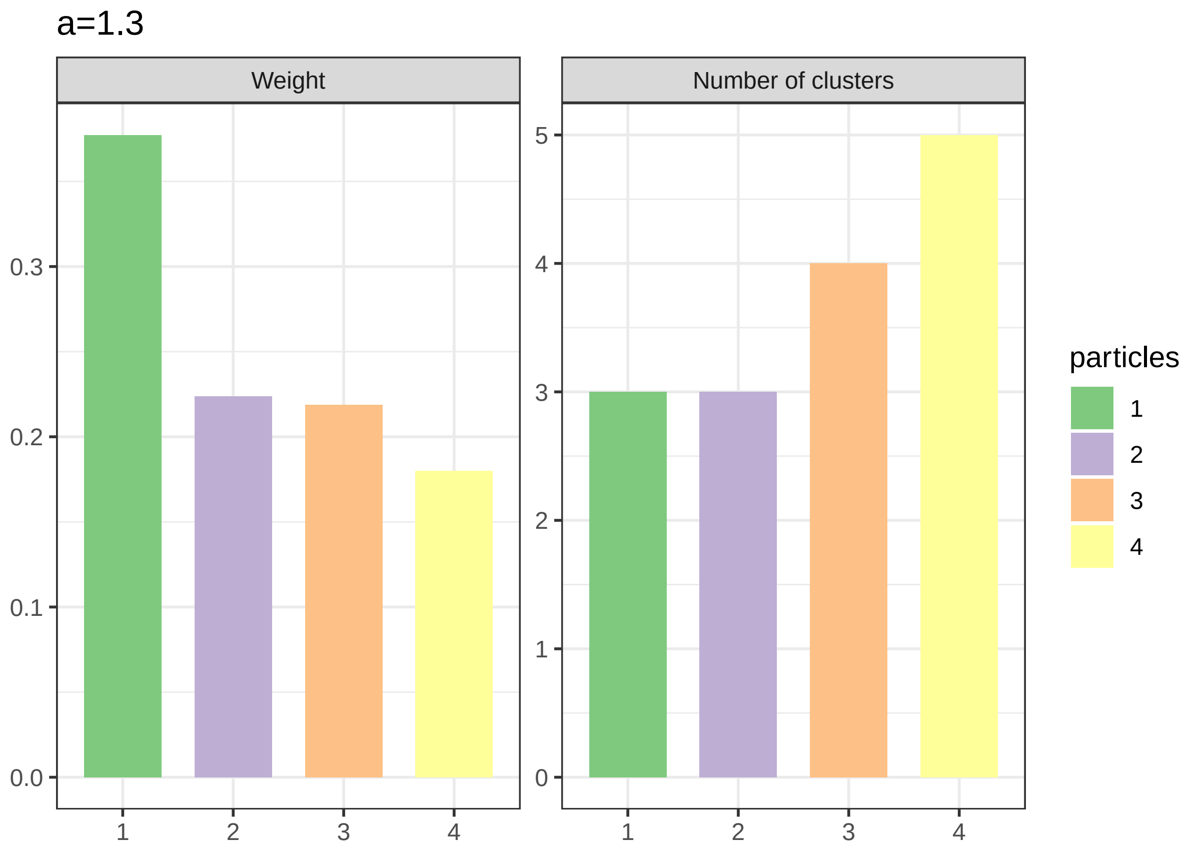


We choose "L=4" as the optimal number of clusters.

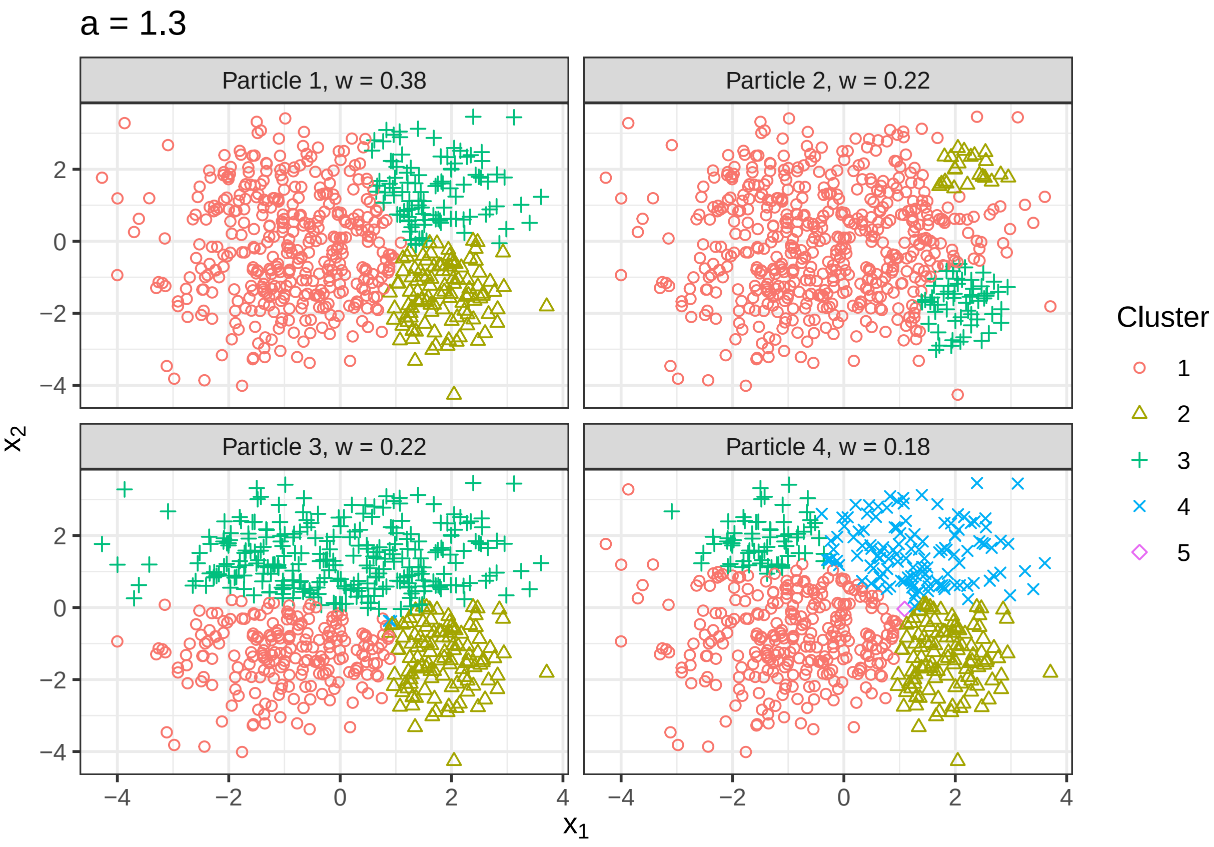
```
L = 4
output_WASABI <- out_elbow$output_list[[L]]
output_WASABI_mb = WASABI_multistart(cls.draw, psm,
                                     multi.start = 25, ncores = 6,
                                     method.init ="++", add_topvi = FALSE,
                                     method="salso", 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, title = "a=1.3")
```



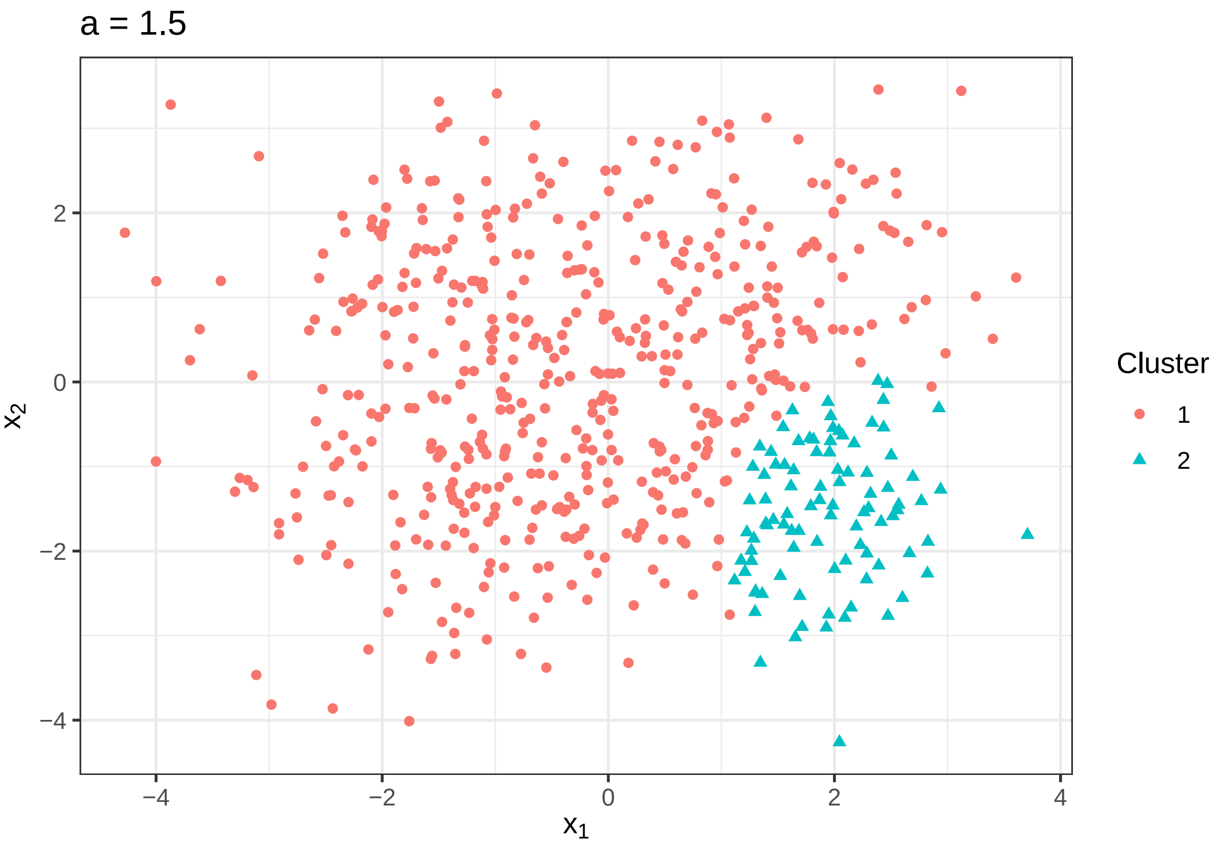
```r
ggscatter_grid2d(output_WASABI, Y, title = "a = 1.3")
```



a = 1.5

```r
table(z_minb5)
```

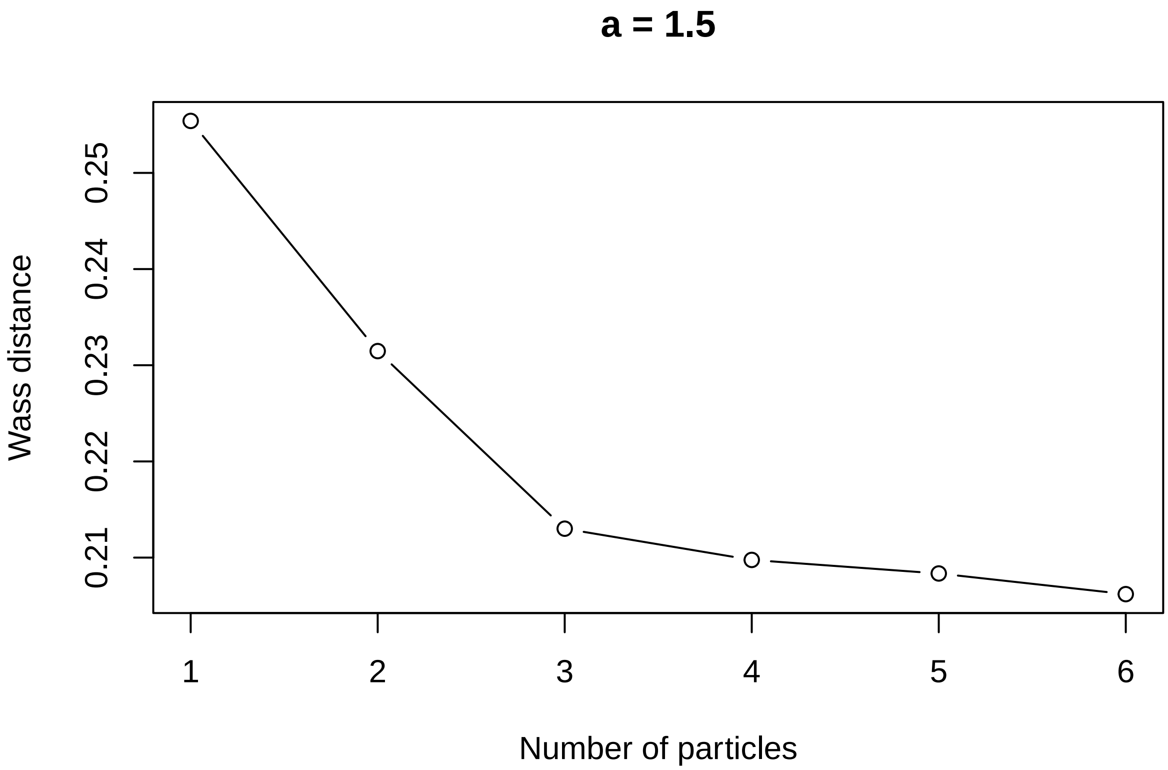
```
## z_minb5
##   1   2
## 506  94
```



```
set.seed(123)
out_elbow <- elbow(cls.draw, L_max = 6, psm = psm,
                   multi.start = 6, method.init = "++",
                   method = "salso", mini.batch = 500, ncores = 6,
                   loss = "Binder", a = 1.5, maxNClusters = 10)
```

```
## Completed  1 / 6
## Completed  2 / 6
## Completed  3 / 6
## Completed  4 / 6
## Completed  5 / 6
## Completed  6 / 6
```

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

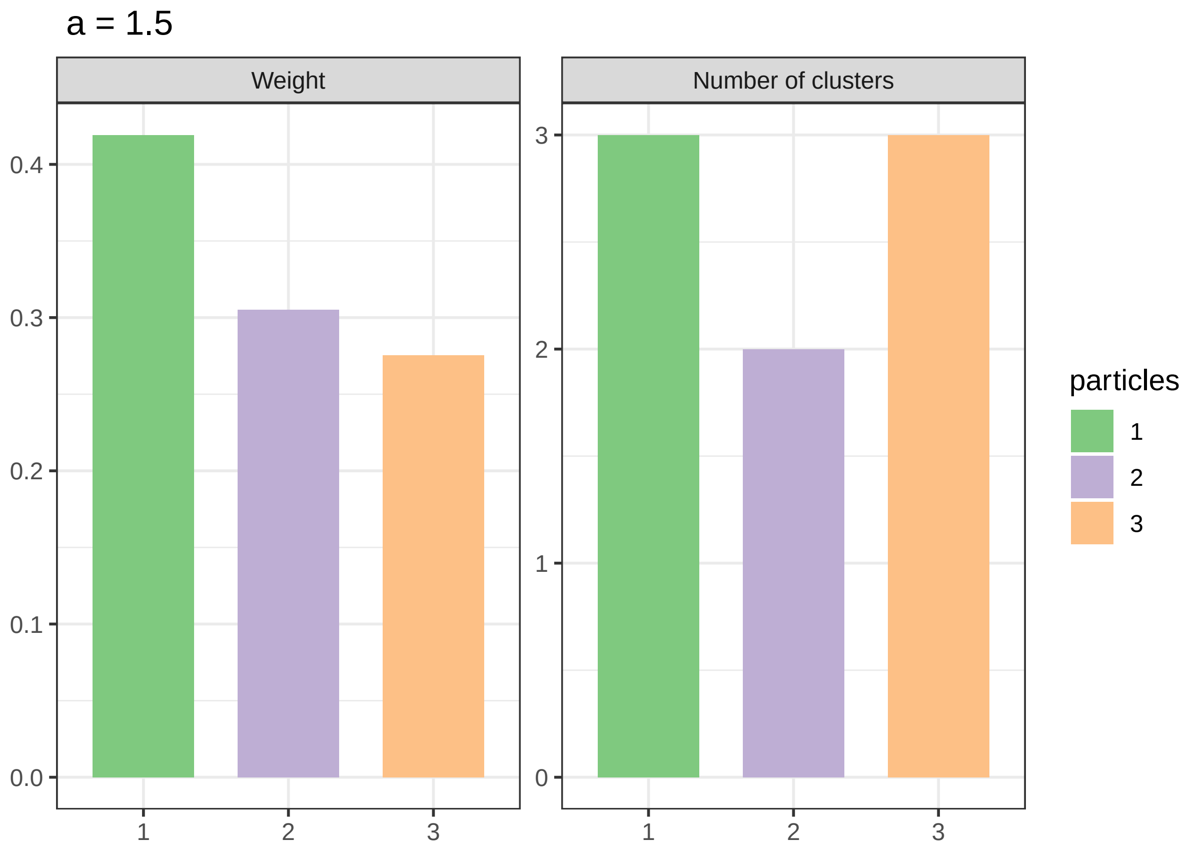


We choose “L=3” as the optimal number of clusters.

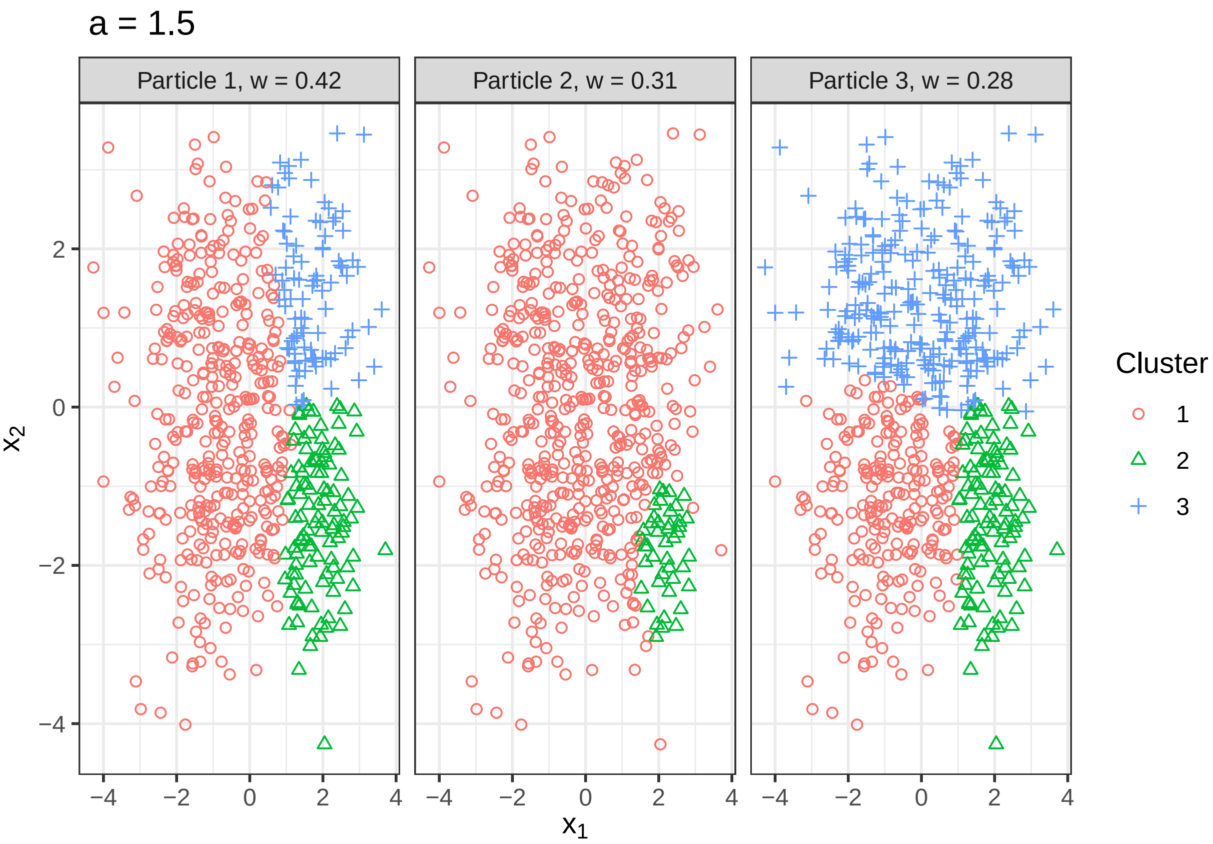
```
L = 3
output_WASABI <- out_elbow$output_list[[L]]
output_WASABI_mb = WASABI_multistart(cls.draw, psm,
                                     multi.start = 25, ncores = 6,
                                     method.init ="++", add_topvi = FALSE,
                                     method="salso", L=L,
                                     mini.batch = 500,
                                     max.iter= 10, extra.iter = 5,
                                     suppress.comment=TRUE,
                                     swap_countone = TRUE,
                                     seed = 54321, loss = "Binder", a = 1.5,
                                     maxNClusters = 10)

if(output_WASABI_mb$wass.dist < output_WASABI$wass.dist){
  output_WASABI <- output_WASABI_mb
}

ggsummary(output_WASABI, title = " a = 1.5 ")
```



```
ggscatter_grid2d(output_WASABI, Y, title = " a = 1.5" )
```


a = 1.5
Particle 1, w = 0.42
Particle 2, w = 0.31
Particle 3, w = 0.28
x_2
2
0
−2
−4
−4
−2
0
2
4
x_1
Cluster
1
2
3