

## Version 4: 3 clusters, 5 samples, 50 variants– linear

### Simulate data

```
I <- 50
K <- 3
S <- 5

set.seed(123)

pi <- c(0.2, 0.5, 0.3)
z <- sample(1:K, size = I, replace = T, prob = pi)

w <- matrix(c(0.9, 0.95, 0.9, 0.8, 0.9,
              0.7, 0.8, 0.85, 0.78, 0.7,
              0.7, 0.2, 0.83, 0.76, 0.5),
            byrow=T,
            nrow=K, ncol=S)
colnames(w) <- paste0("sample", 1:S)

tcn <- matrix(2, nrow=I, ncol=S)
m <- matrix(rep(sample(1:2, size = I, replace = T), S),
            nrow=I, ncol=S)
W <- w[z, ]
calcTheta <- function(m, tcn, w) {
  (m * w) / (tcn * w + 2*(1-w))
}
theta <- calcTheta(m, tcn, W)

n <- replicate(S, rpois(I, 100))
y <- matrix(NA, nrow=I, ncol=S)
for (i in 1:I) {
  for (s in 1:S) {
    y[i, s] <- rbinom(1, n[i, s], theta[i,s])
  }
}

test.data <- list("I" = I, "S" = S, "K" = K,
                 "y" = y, "n" = n,
                 "m" = m, "tcn" = tcn)
```

### Visualize densities of simulated data

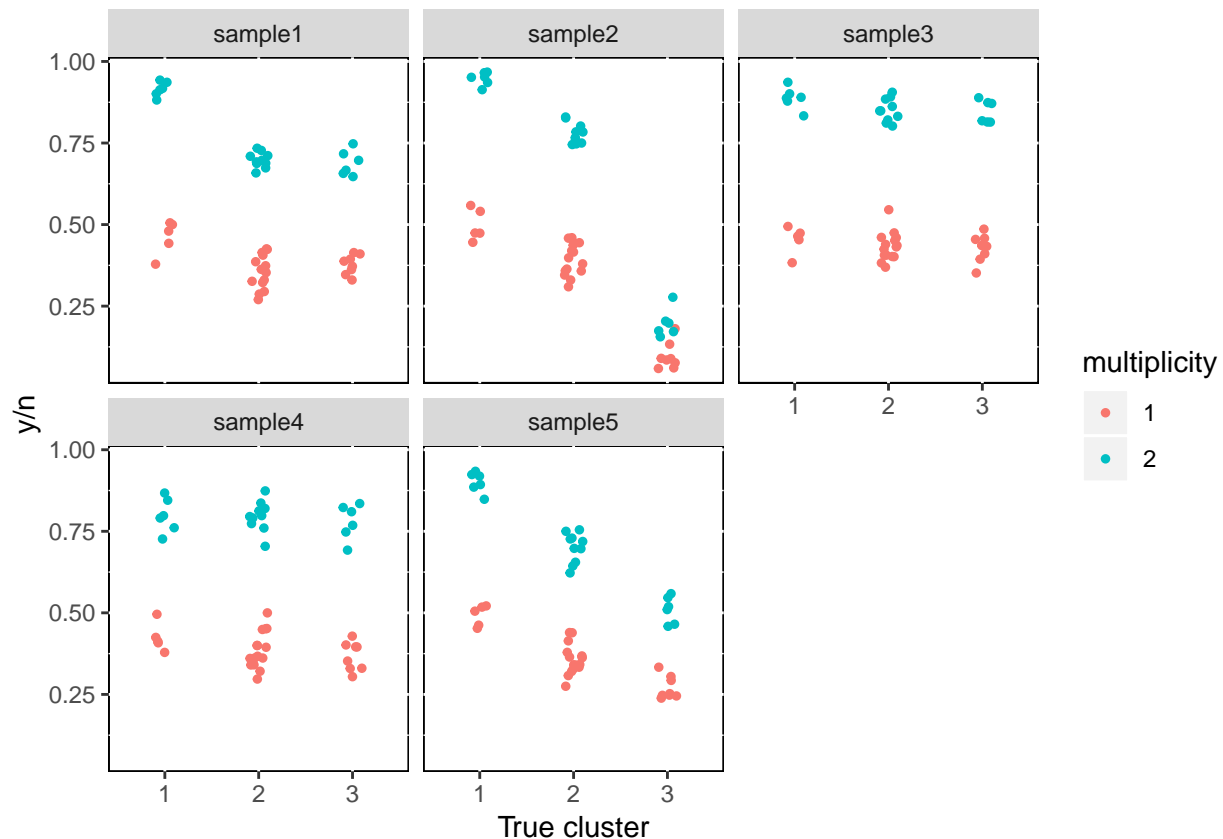
Clustering is by  $\omega$

```
p <- test.data$y/test.data$n
colnames(w) <- colnames(p) <- paste0("sample", 1:S)
colnames(m) <- colnames(p)
m2 <- m %>%
  as_tibble() %>%
```

```

mutate(z=factor(z),
       variant_index=seq_len(nrow(.))) %>%
gather("sample", "multiplicity", -c(z, variant_index)) %>%
mutate(sample=factor(sample, levels=colnames(p)))
p2 <- p %>%
  as_tibble() %>%
  mutate(z=factor(z),
         variant_index=seq_len(nrow(.))) %>%
  gather("sample", "fraction", -c(z, variant_index)) %>%
  mutate(sample=factor(sample, levels=colnames(p))) %>%
  left_join(m2, by=c("z", "variant_index", "sample")) %>%
  mutate(multiplicity=factor(multiplicity))
## what the data looks like by sample
ggplot(p2, aes(z, fraction)) +
  geom_jitter(width=0.1, size=1, aes(color=multiplicity)) +
  facet_wrap(~sample) +
  ylab("y/n") +
  xlab("True cluster") +
  theme(panel.background=element_rect(fill="white", color="black"))

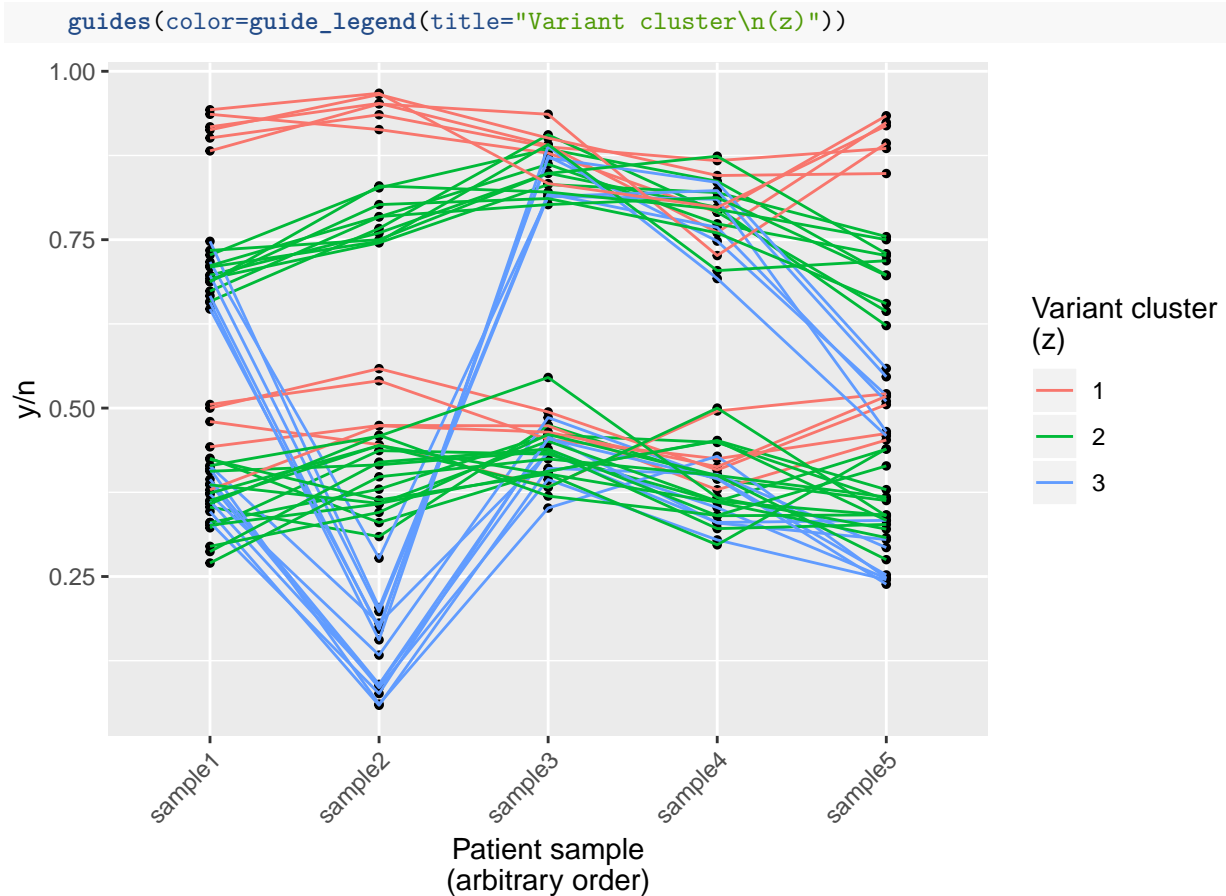
```



```

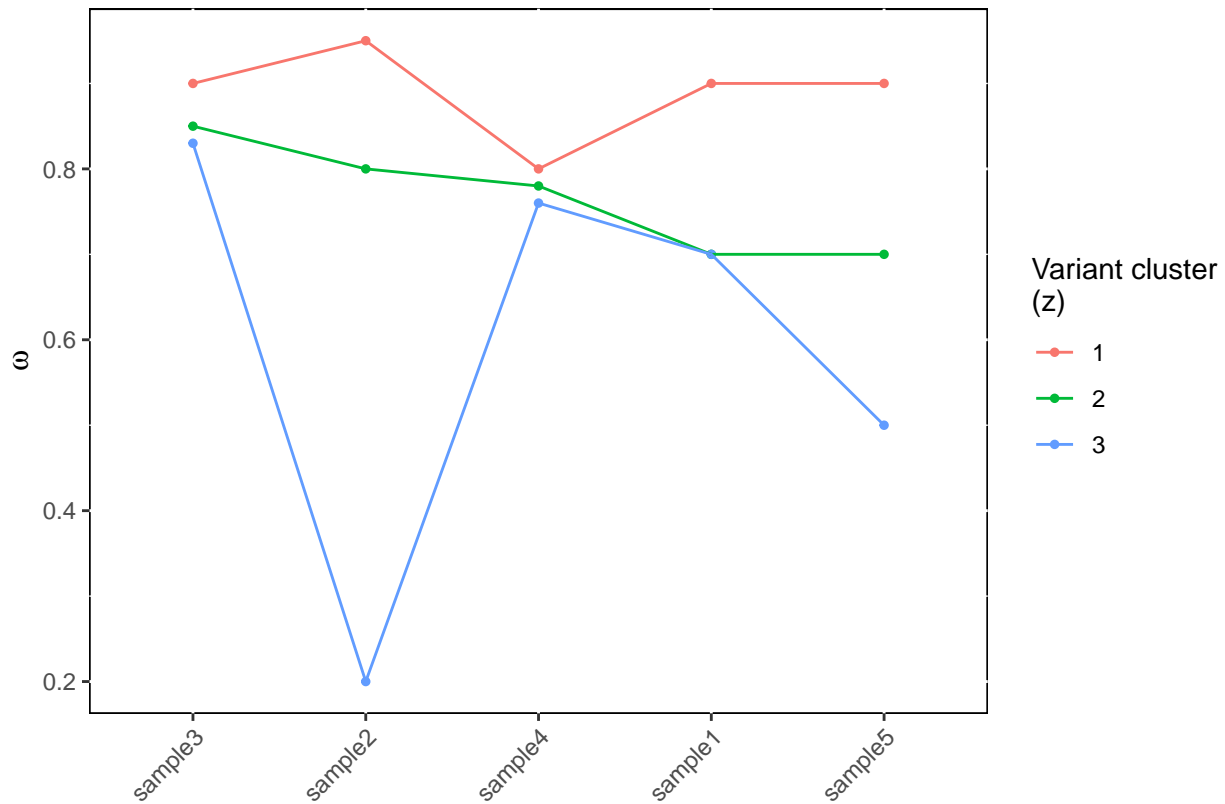
## by variant
ggplot(p2, aes(sample, fraction, group=variant_index)) +
  geom_point(size=1) +
  geom_line(aes(color=z)) +
  ylab("y/n") +
  xlab("Patient sample\n(arbitrary order)") +
  theme(axis.text.x=element_text(angle=45, hjust=1)) +

```



```
##
## Cluster means and variances
##
meds <- apply(W, 2, median)
slevels <- colnames(w)[order(meds, decreasing=TRUE)]
w2 <- w %>%
  as_tibble() %>%
  mutate(z=factor(1:K)) %>%
  ## variant_index=seq_len(nrow(.)) %>%
  gather("sample", "omega", -z) %>%
  mutate(sample=factor(sample, levels=slevels))
##fig1 <- ggplot(w2, aes(z, omega)) +
##   geom_jitter(width=0.1, size=1) +
##   facet_wrap(~sample) +
##   ylab(expression(omega)) +
##   xlab("True cluster") +
##   theme(panel.background=element_rect(fill="white", color="black"))
fig2 <- ggplot(w2, aes(sample, omega, group=z)) +
  geom_point(size=1, aes(color=z)) +
  geom_line(aes(color=z)) +
  ylab(expression(omega)) +
  xlab("") +
  theme(axis.text.x=element_text(angle=45, hjust=1),
        panel.background=element_rect(fill="white", color="black"),
        legend.key=element_rect(fill="white", color="white")) +
```

```
guides(color=guide_legend(title="Variant cluster\n(z)"))
fig2
```



## JAGS

```
jags.file <- file.path(models.dir, "w.jags")
inits <- list(".RNG.name" = "base::Wichmann-Hill",
              ".RNG.seed" = 123)
jags.m <- jags.model(jags.file, test.data,
                    n.chains = 1,
                    inits = inits,
                    n.adapt = 1000)

## Compiling model graph
##   Resolving undeclared variables
##   Allocating nodes
## Graph information:
##   Observed stochastic nodes: 250
##   Unobserved stochastic nodes: 316
##   Total graph size: 3963
##
## Initializing model

params <- c("z", "w", "ystar")
samps <- coda.samples(jags.m, params, n.iter=10000, thin=7)
jags_df <- ggs(samps)
#ggs_traceplot(jags_df, family="w")
```

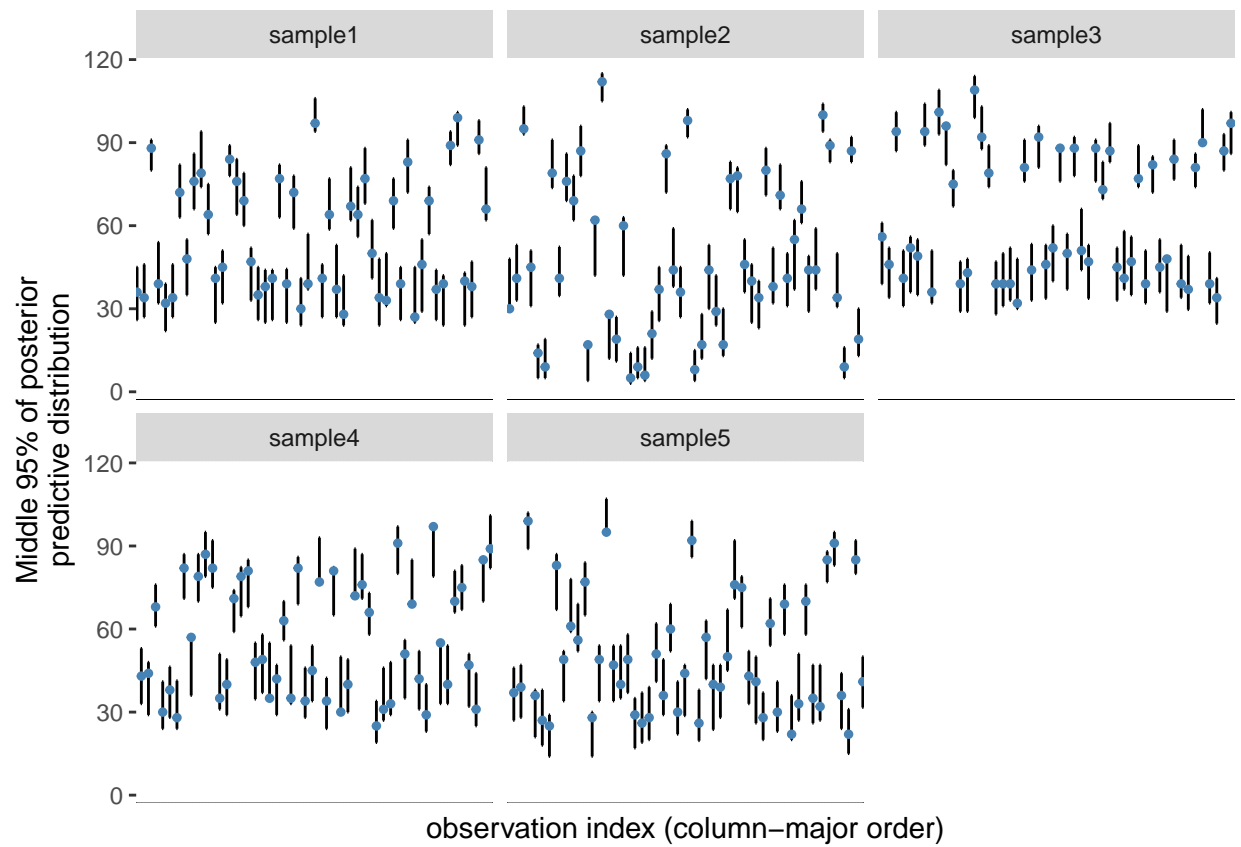
```

s <- summary(samps)
#effectiveSize(samps)
pdf(file.path(trace.dir, paste0(runName, "_trace.pdf")))
plot(samps)
dev.off()

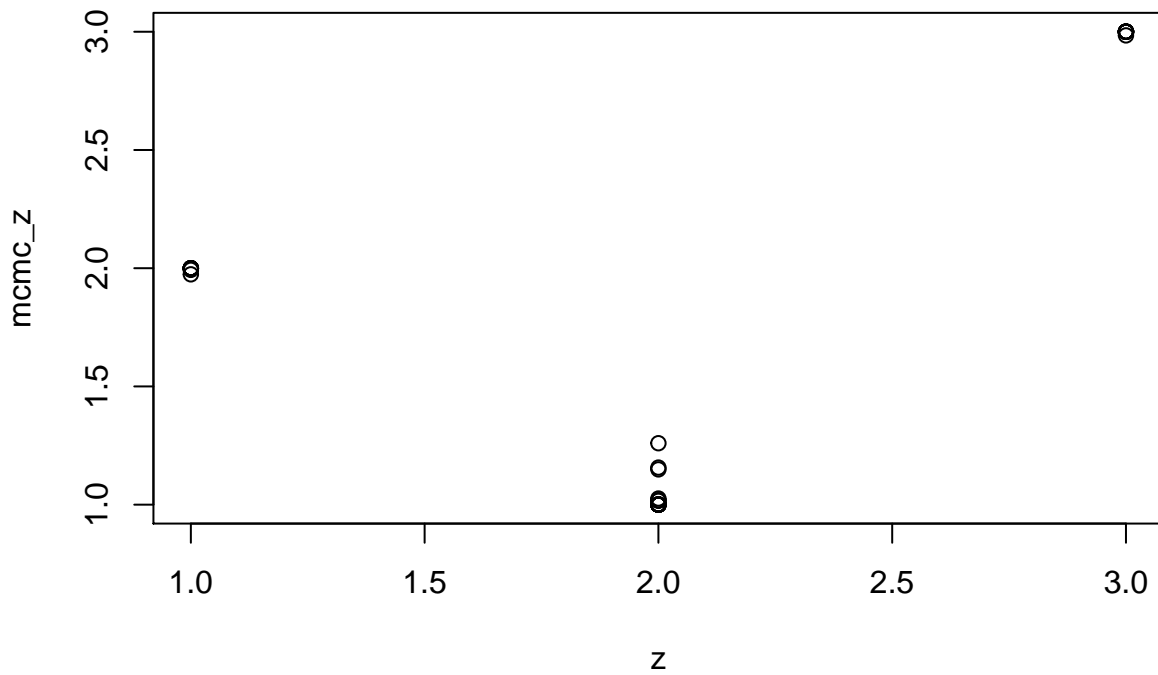
## pdf
## 2

##
## 50 mutations x 10 samples
chains <- do.call(rbind, samps)
ystar <- chains[, grep("ystar", colnames(chains))]
## each row of MCMC is in column-major order
orig.order <- tibble(statistic=colnames(ystar))
ppd.summaries <- ystar %>%
  as_tibble() %>%
  gather("statistic", "value") %>%
  group_by(statistic) %>%
  summarize(mean=mean(value),
            q1=quantile(value, 0.025),
            q3=quantile(value, 0.975))
ppd.summaries2 <- left_join(orig.order,
                           ppd.summaries, by="statistic") %>%
  mutate(observed=as.numeric(test.data$y)) %>%
  mutate(sample=paste0("sample", rep(1:S, each=I)),
         variant=rep(1:I, S))
ggplot(ppd.summaries2, aes(x=statistic, y=mean,
                          ymin=q1,
                          ymax=q3)) +
  geom_errorbar() +
  geom_point(aes(x=statistic, y=observed),
            size=1, color="steelblue") +
  theme(axis.text.x=element_blank(),
        axis.ticks.x=element_blank(),
        panel.background=element_rect(fill="white",
                                       color="black")) +
  ylab("Middle 95% of posterior\npredictive distribution") +
  xlab("observation index (column-major order)") +
  facet_wrap(~sample)

```



## Plots



z	mcmc_z
2	1
1	2
3	3

