TARGET_GMM_STAN

Graham Gumbert

2024-12-20

```
data <- read.table(paste0(params$cloud_path, params$beta_vals_file),
                   header = TRUE, sep = "\t")

site_ids <- data[[1]]
tumor_data <- data[,-1]

num_tumors <- ncol(tumor_data)
num_sites <- nrow(tumor_data)

cat("Number of sites:", num_sites, "\n")
```

```
## Number of sites: 1000
```

```
cat("Number of tumors:", num_tumors, "\n")
```

```
## Number of tumors: 151
```

```
stan_code <- "
data {
  int<lower=1> N_data;
  array[N_data] real y;
}

parameters {
  simplex[3] psi; // Mixture weights sum to 1
  real<lower=0,upper=0.5> X; // left peak position
  real<lower=1> N; // effective number of alleles
}

transformed parameters {
  // Means of the three Gaussian peaks
  real mu_left = X;
  real mu_right = 1 - X;
  real mu_middle = 0.5;

  // Variances (p*(1-p)/N) for each peak
  real var_left = X*(1-X)/N;
  real var_right = X*(1-X)/N;
  real var_middle = 0.25/N; // 0.5*(1-0.5)=0.25
```

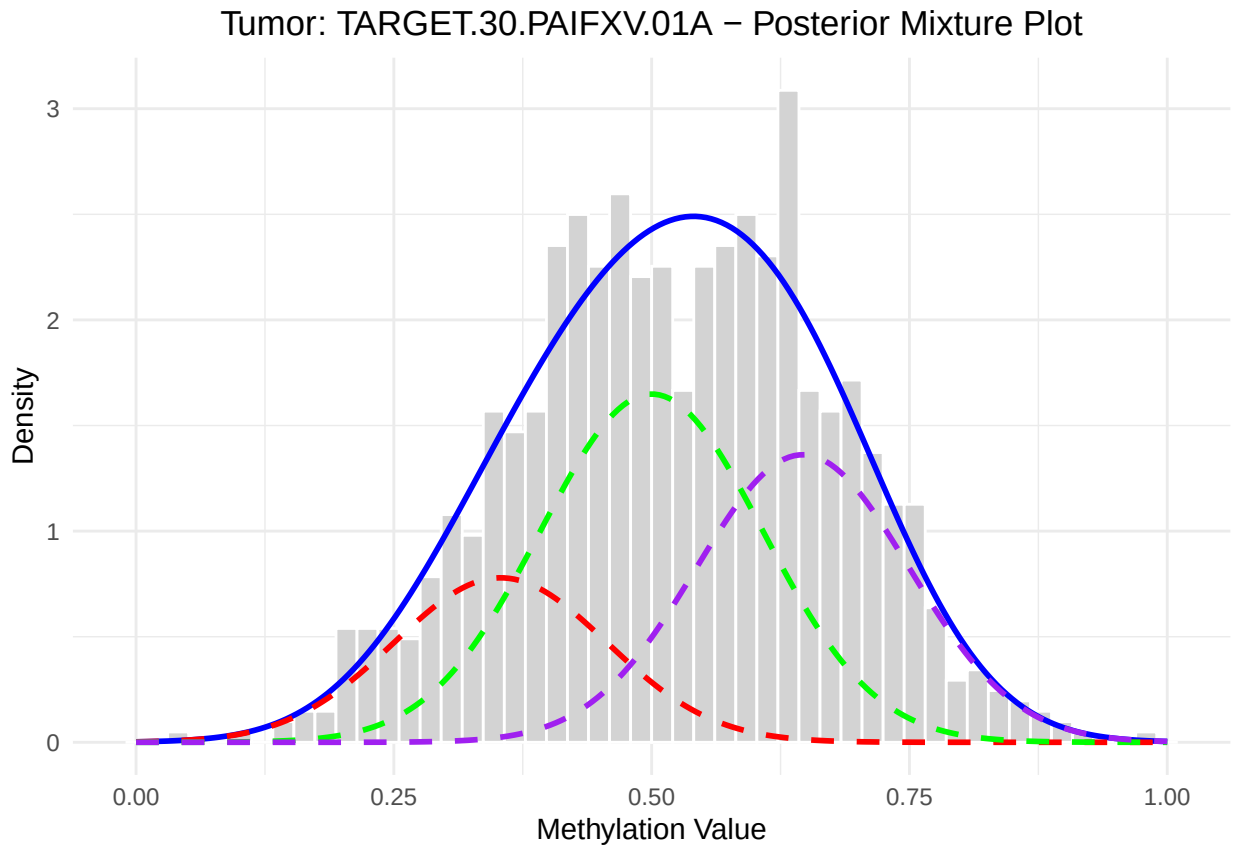
```
  // Standard deviations
  real sigma_left = sqrt(var_left);
  real sigma_right = sqrt(var_right);
  real sigma_middle = sqrt(var_middle);
}

model {
  psi ~ dirichlet([5.0, 10.0, 5.0]); // quarter half quarter weighting with tuning for variance
  X ~ beta(1,4);
  N ~ gamma(2, 0.1);

  for (n in 1:N_data) {
    array[3] real lps;
    lps[1] = log(psi[1]) + normal_lpdf(y[n] | mu_left, sigma_left);
    lps[2] = log(psi[2]) + normal_lpdf(y[n] | mu_middle, sigma_middle);
    lps[3] = log(psi[3]) + normal_lpdf(y[n] | mu_right, sigma_right);

    target += log_sum_exp(lps);
  }
}

generated quantities {
  real phi = -log1m(2 * X) / 2;
}
"

sm <- stan_model(model_code = stan_code)
```

```
## Running /Library/Frameworks/R.framework/Resources/bin/R CMD SHLIB foo.c
## using C compiler: 'Apple clang version 16.0.0 (clang-1600.0.26.6)'
## using SDK: 'MacOSX15.2.sdk'
## clang -arch arm64 -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG   -I"/Library/Frame
## In file included from <built-in>:1:
## In file included from /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library/StanHeade
## In file included from /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library/RcppEigen
## In file included from /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library/RcppEigen
## /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/library/RcppEigen/include/Eigen/src/Cor
##   679 | #include <cmath>
##       |          ^~~~~~~
## 1 error generated.
## make: *** [foo.o] Error 1
```

```
tumor_params <- data.frame(
  Tumor = names(results_list),
  psi_1 = NA,
  psi_2 = NA,
  psi_3 = NA,
  X     = NA,
  N     = NA,
  phi   = NA,
  stringsAsFactors = FALSE
)
```

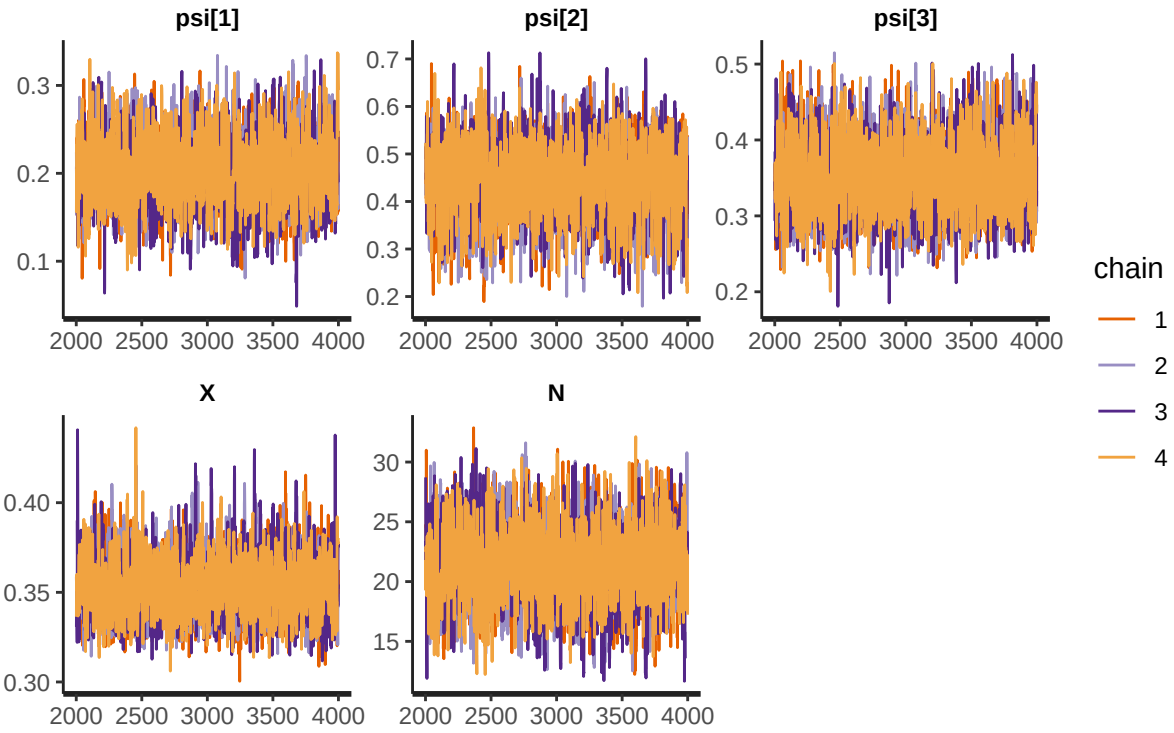
```r
for (i in 1:length(results_list)) {
  fit <- results_list[[i]]
  posterior_samples <- rstan::extract(fit)

  # Posterior means
  psi_1_mean <- mean(posterior_samples$psi[,1])
  psi_2_mean <- mean(posterior_samples$psi[,2])
  psi_3_mean <- mean(posterior_samples$psi[,3])
  X_mean     <- mean(posterior_samples$X)
  N_mean     <- mean(posterior_samples$N)
  phi_mean     <- mean(posterior_samples$phi)

  tumor_params[i, "psi_1"] <- psi_1_mean
  tumor_params[i, "psi_2"] <- psi_2_mean
  tumor_params[i, "psi_3"] <- psi_3_mean
  tumor_params[i, "X"]     <- X_mean
  tumor_params[i, "N"]     <- N_mean
  tumor_params[i, "phi"]   <- phi_mean
}

tumor_params
```

```
##                  Tumor     psi_1     psi_2     psi_3         X        N
## 1 TARGET.30.PAIFXV.01A 0.2015072 0.4462108 0.3522820 0.3534102 21.44872
## 2 TARGET.30.PAISNS.01A 0.2280895 0.4297337 0.3421767 0.2384317 21.12472
## 3 TARGET.30.PAITCI.01A 0.2205359 0.3516320 0.4278321 0.2868129 22.44453
##         phi
## 1 0.6168009
## 2 0.3241160
## 3 0.4265988
```

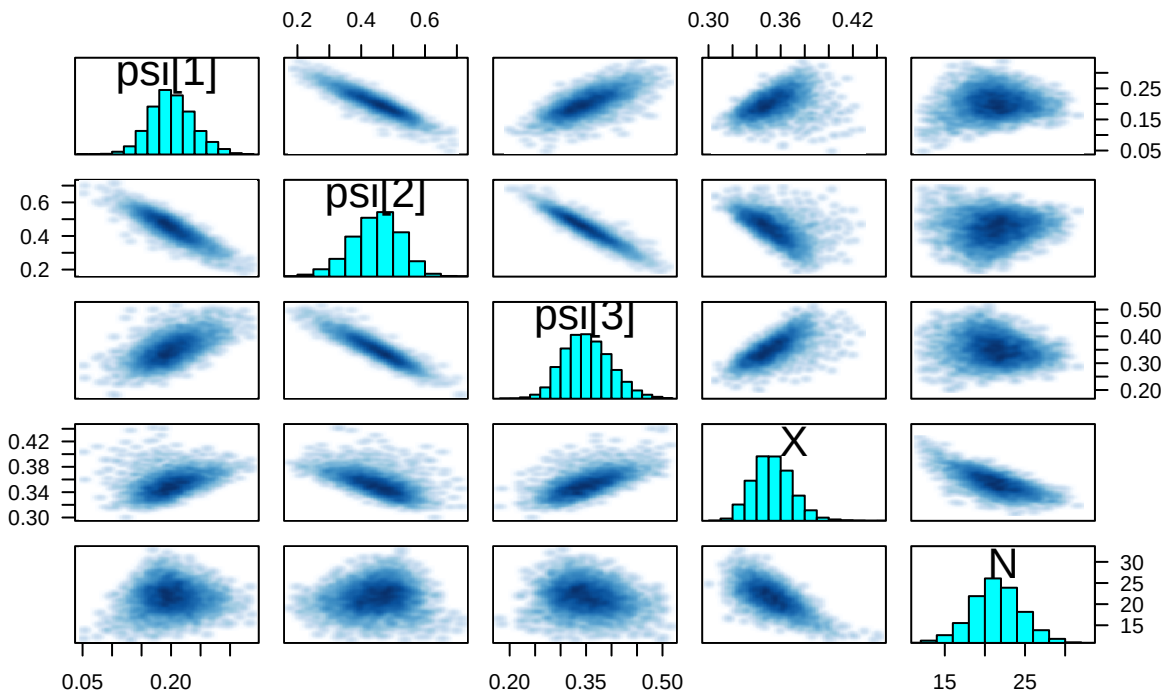



```
## ### Tumor: TARGET.30.PAIFXV.01A - Posterior Summary (Stan Fit)
## Inference for Stan model: anon_model.
## 4 chains, each with iter=4000; warmup=2000; thin=1;
## post-warmup draws per chain=2000, total post-warmup draws=8000.
##
##         mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## psi[1]  0.20    0.00 0.04  0.13  0.18  0.20  0.23  0.28  2001    1
## psi[2]  0.45    0.00 0.08  0.29  0.40  0.45  0.50  0.58  2038    1
## psi[3]  0.35    0.00 0.05  0.27  0.32  0.35  0.38  0.45  2365    1
## X       0.35    0.00 0.02  0.32  0.34  0.35  0.36  0.39  1992    1
## N      21.45    0.07 3.08 15.24 19.43 21.44 23.49 27.57  2202    1
## phi     0.62    0.00 0.06  0.52  0.58  0.61  0.65  0.75  1943    1
##
## Samples were drawn using NUTS(diag_e) at Sun Feb 23 21:39:37 2025.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

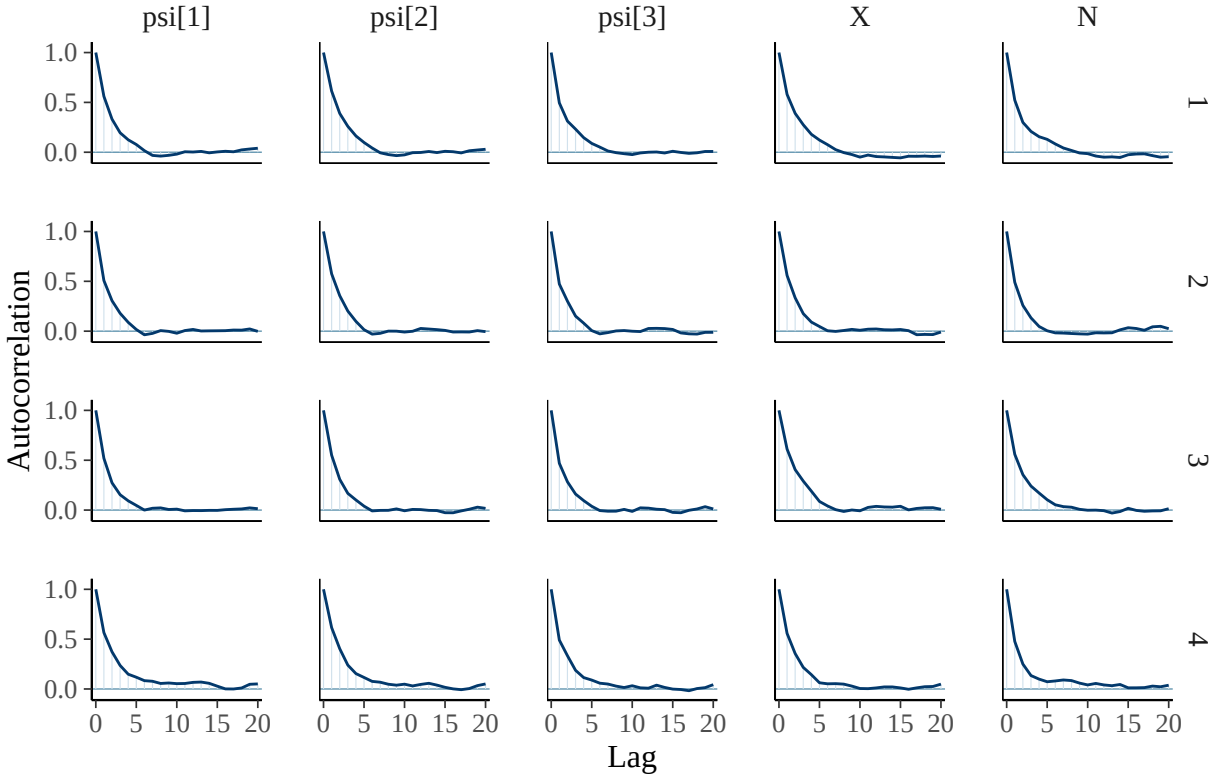
Traceplot for tumor: TARGET.30.PAIFXV.01A

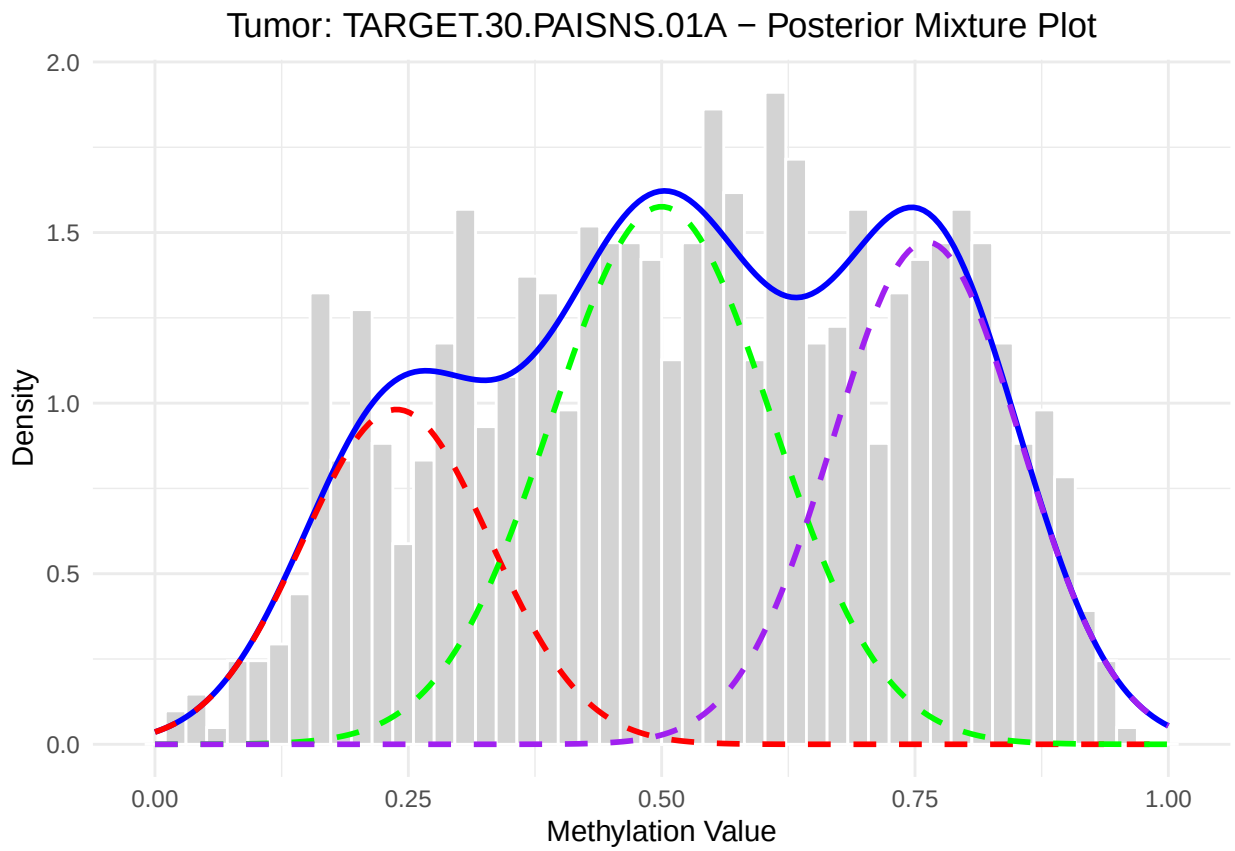


Pairs Plot for tumor: TARGET.30.PAIFXV.01A



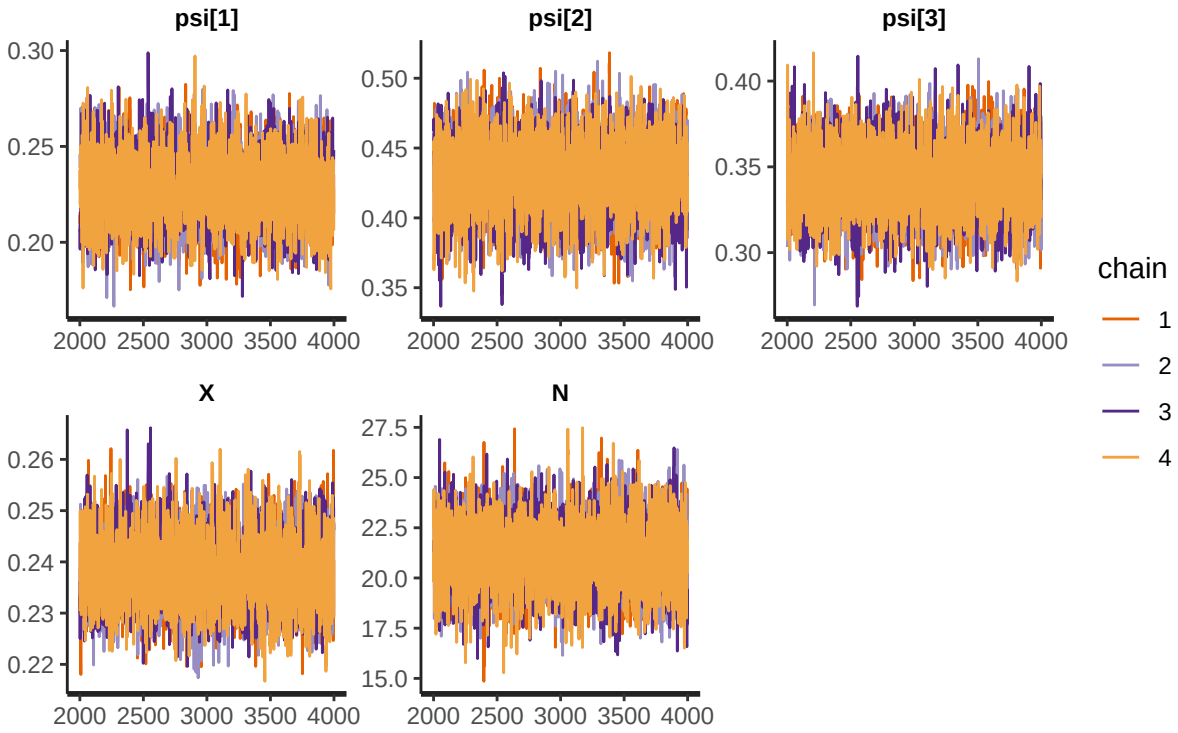
Autocorrelation Plot for Tumor: TARGET.30.PAIFXV.01A



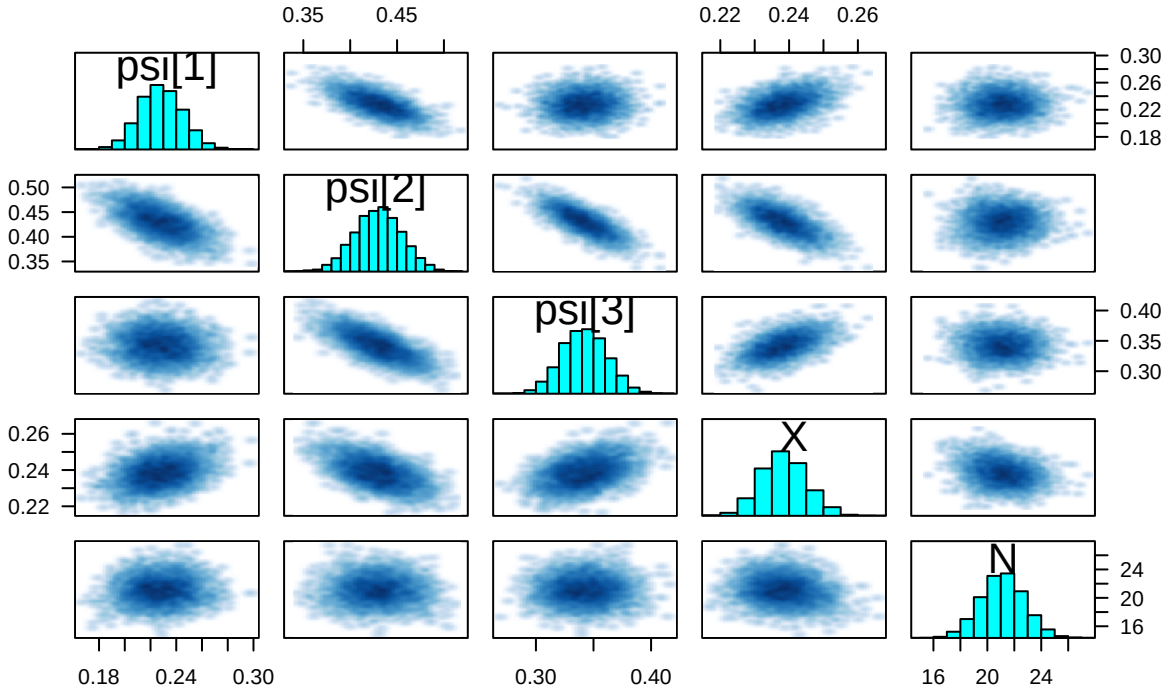



```
## ### Tumor: TARGET.30.PAISNS.01A - Posterior Summary (Stan Fit)
## Inference for Stan model: anon_model.
## 4 chains, each with iter=4000; warmup=2000; thin=1;
## post-warmup draws per chain=2000, total post-warmup draws=8000.
##
##         mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## psi[1]  0.23    0.00 0.02  0.20  0.22  0.23  0.24  0.26  4941    1
## psi[2]  0.43    0.00 0.03  0.38  0.41  0.43  0.45  0.48  4116    1
## psi[3]  0.34    0.00 0.02  0.30  0.33  0.34  0.36  0.38  5394    1
## X       0.24    0.00 0.01  0.23  0.23  0.24  0.24  0.25  4043    1
## N      21.12    0.02 1.63 17.94 20.02 21.11 22.21 24.32  4961    1
## phi     0.32    0.00 0.01  0.30  0.32  0.32  0.33  0.35  4024    1
##
## Samples were drawn using NUTS(diag_e) at Sun Feb 23 21:39:48 2025.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

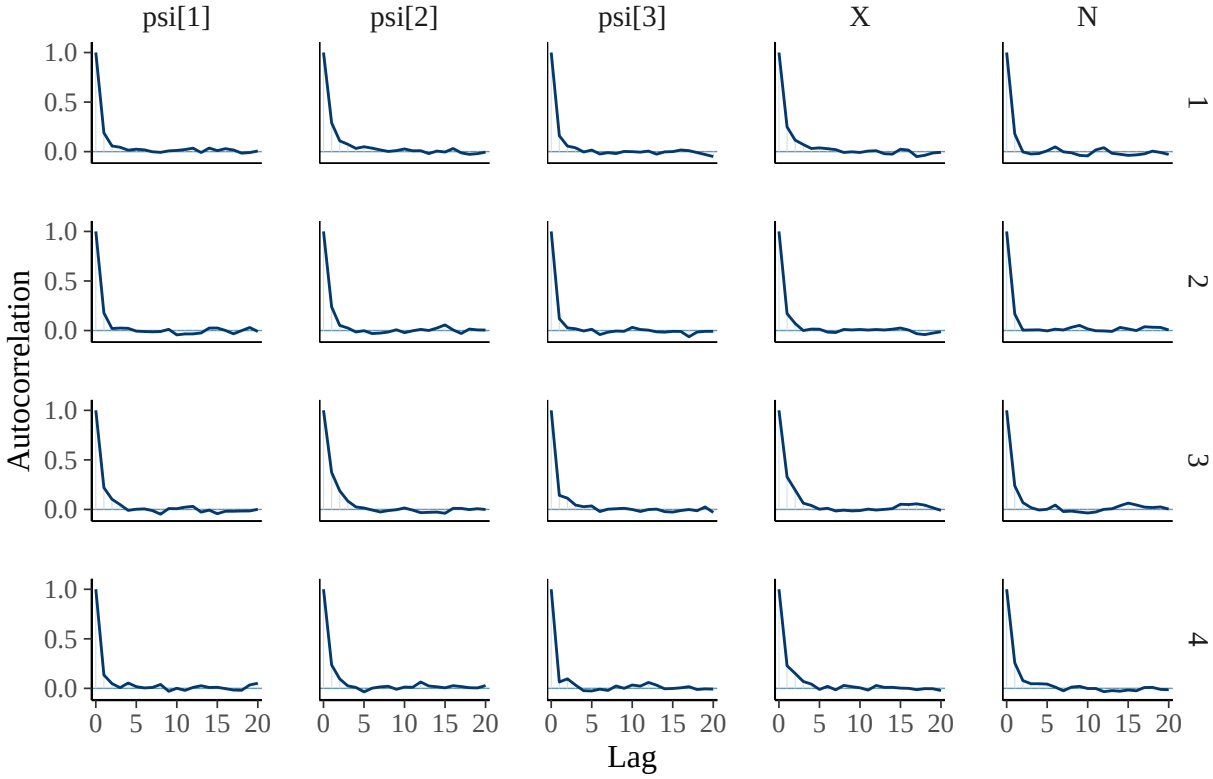
Traceplot for tumor: TARGET.30.PAISNS.01A

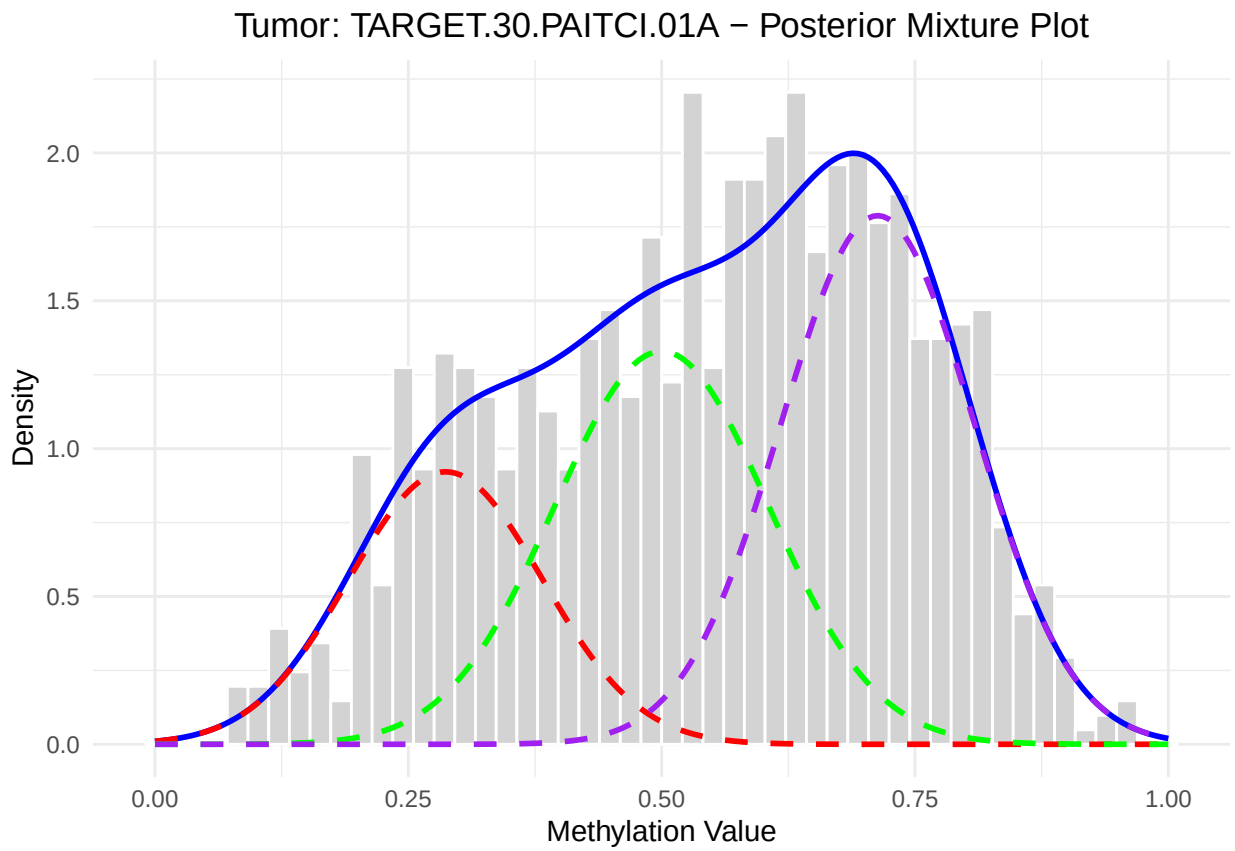


Pairs Plot for tumor: TARGET.30.PAISNS.01A



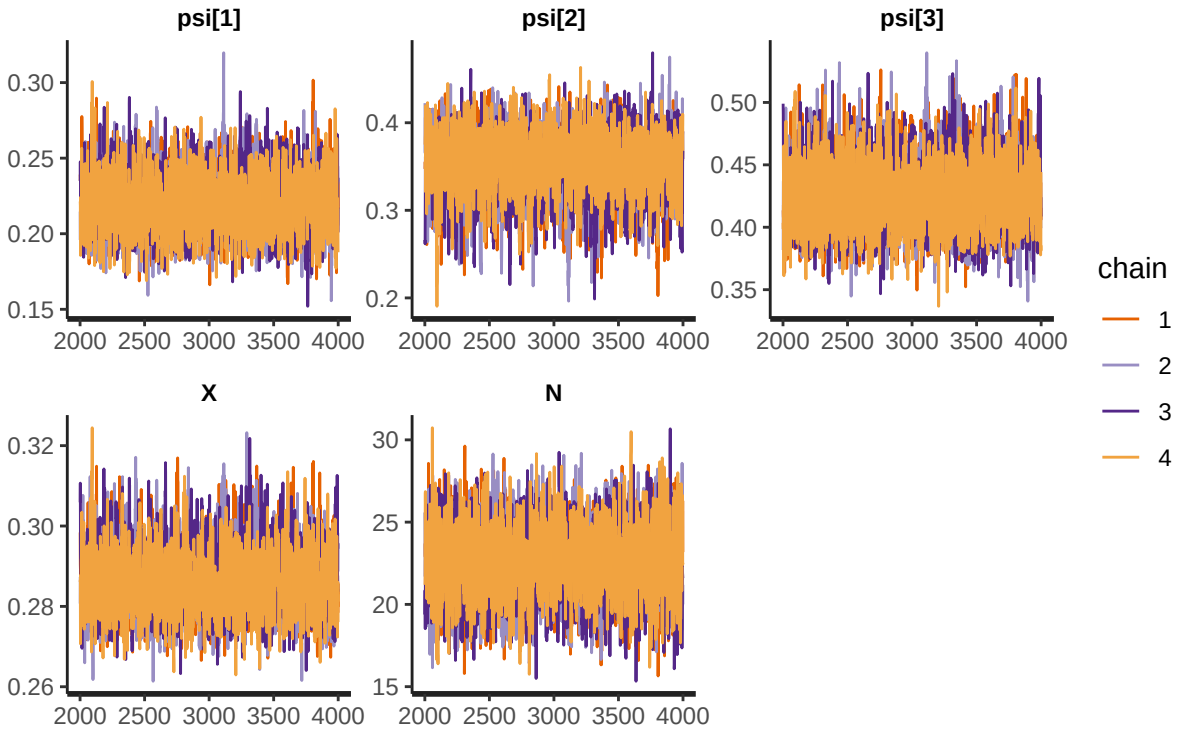
Autocorrelation Plot for Tumor: TARGET.30.PAISNS.01A



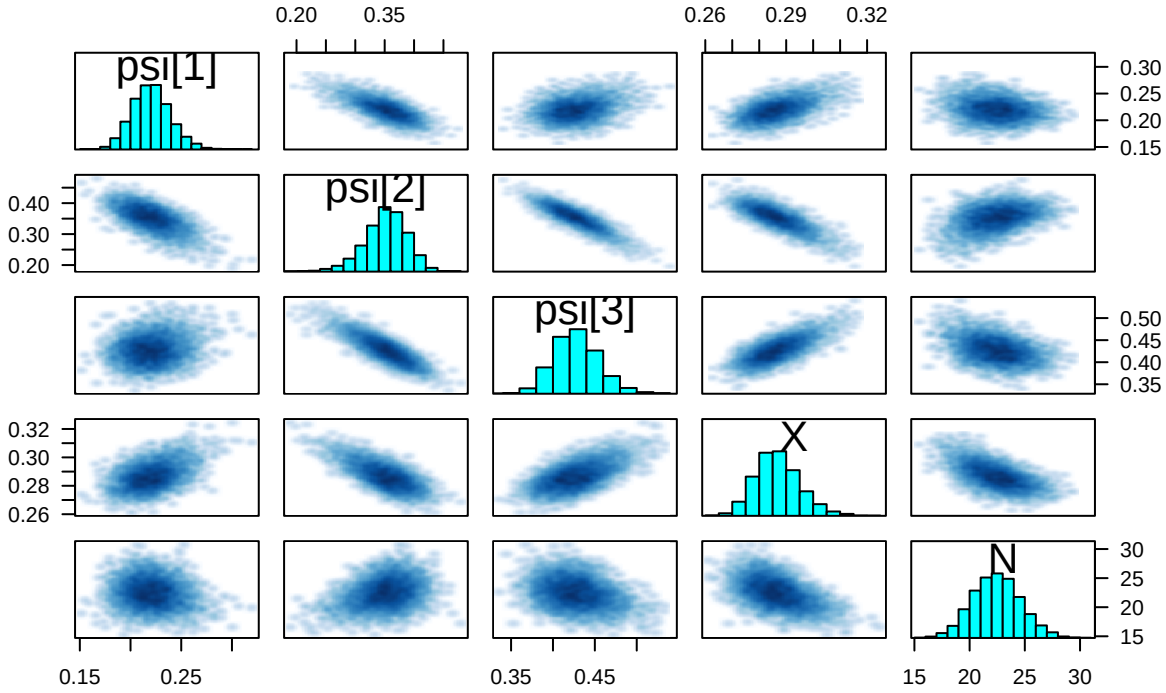



```
## ### Tumor: TARGET.30.PAITCI.01A - Posterior Summary (Stan Fit)
## Inference for Stan model: anon_model.
## 4 chains, each with iter=4000; warmup=2000; thin=1;
## post-warmup draws per chain=2000, total post-warmup draws=8000.
##
##         mean se_mean   sd  2.5%   25%   50%   75% 97.5% n_eff Rhat
## psi[1]  0.22    0.00 0.02  0.19  0.21  0.22  0.23  0.26  2773    1
## psi[2]  0.35    0.00 0.04  0.27  0.33  0.35  0.38  0.42  2183    1
## psi[3]  0.43    0.00 0.03  0.38  0.41  0.43  0.44  0.48  2929    1
## X       0.29    0.00 0.01  0.27  0.28  0.29  0.29  0.30  2221    1
## N      22.44    0.04 2.17 18.28 20.93 22.42 23.88 26.82  2910    1
## phi     0.43    0.00 0.02  0.39  0.41  0.42  0.44  0.47  2180    1
##
## Samples were drawn using NUTS(diag_e) at Sun Feb 23 21:40:02 2025.
## For each parameter, n_eff is a crude measure of effective sample size,
## and Rhat is the potential scale reduction factor on split chains (at
## convergence, Rhat=1).
```

Traceplot for tumor: TARGET.30.PAITCI.01A



Pairs Plot for tumor: TARGET.30.PAITCI.01A



Autocorrelation Plot for Tumor: TARGET.30.PAITCI.01A