Simulating Drop-seq scRNA count data

```
# Install scSimu and related packages
if (!requireNamespace("scSimu", quietly = TRUE)){
  devtools::install_github("xs222/scSimu")
}
if (!requireNamespace("glmGamPoi", quietly = TRUE)){
  if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")
  BiocManager::install("glmGamPoi")
}
if (!requireNamespace("CSCORE", quietly = TRUE)){
  devtools::install_github("ChangSuBiostats/CS-CORE")
}


library(scSimu)
```

Load data

Here we utilize the T cells from 2,700 PBMC from Seurat.

```
SeuratData::InstallData("pbmc3k")
#> Warning: The following packages are already installed and will not be
#> reinstalled: pbmc3k
```

```
library(Seurat)
#> Loading required package: SeuratObject
#> Loading required package: sp
#>
#> Attaching package: 'SeuratObject'
#> The following object is masked from 'package:base':
#>
#>     intersect
```

```
library(SeuratData)
#> -- Installed datasets ------------------------------- SeuratData v0.2.2.9001 --
#> v pbmc3k       3.1.4                     v pbmcMultiome 0.1.4
#> ---------------------------------------- Key -----------------------------------------
#> v Dataset loaded successfully
#> > Dataset built with a newer version of Seurat than installed
#> (?) Unknown version of Seurat installed
```

```
library(ggplot2)
library(patchwork)
pbmc3k.final <- LoadData("pbmc3k", type = "pbmc3k.final")
#> Validating object structure
#> Updating object slots
#> Ensuring keys are in the proper structure
#> Updating matrix keys for DimReduc 'pca'
#> Updating matrix keys for DimReduc 'umap'
#> Warning: Assay RNA changing from Assay to Assay
#> Warning: Graph RNA_nn changing from Graph to Graph
#> Warning: Graph RNA_snn changing from Graph to Graph
#> Warning: DimReduc pca changing from DimReduc to DimReduc
#> Warning: DimReduc umap changing from DimReduc to DimReduc
#> Ensuring keys are in the proper structure
#> Ensuring feature names don't have underscores or pipes
#> Updating slots in RNA
#> Updating slots in RNA_nn
#> Setting default assay of RNA_nn to RNA
#> Updating slots in RNA_snn
#> Setting default assay of RNA_snn to RNA
#> Updating slots in pca
#> Updating slots in umap
#> Setting umap DimReduc to global
#> Setting assay used for NormalizeData.RNA to RNA
#> Setting assay used for FindVariableFeatures.RNA to RNA
#> Setting assay used for ScaleData.RNA to RNA
#> Setting assay used for RunPCA.RNA to RNA
#> Setting assay used for JackStraw.RNA.pca to RNA
#> No assay information could be found for ScoreJackStraw
#> Warning: Adding a command log without an assay associated with it
#> Setting assay used for FindNeighbors.RNA.pca to RNA
#> No assay information could be found for FindClusters
#> Warning: Adding a command log without an assay associated with it
#> Setting assay used for RunUMAP.RNA.pca to RNA
#> Validating object structure for Assay 'RNA'
#> Validating object structure for Graph 'RNA_nn'
#> Validating object structure for Graph 'RNA_snn'
#> Validating object structure for DimReduc 'pca'
#> Validating object structure for DimReduc 'umap'
#> Object representation is consistent with the most current Seurat version
#> Warning: Assay RNA changing from Assay to Assay5
```

```
pbmc3k.final <- UpdateSeuratObject(pbmc3k.final)
#> Validating object structure
#> Updating object slots
#> Ensuring keys are in the proper structure
#> Updating matrix keys for DimReduc 'pca'
#> Updating matrix keys for DimReduc 'umap'
#> Ensuring keys are in the proper structure
#> Ensuring feature names don't have underscores or pipes
#> Updating slots in RNA
#> Updating slots in RNA_nn
#> Setting default assay of RNA_nn to RNA
```

```
#> Updating slots in RNA_snn
#> Setting default assay of RNA_snn to RNA
#> Updating slots in pca
#> Updating slots in umap
#> Setting umap DimReduc to global
#> Setting assay used for NormalizeData.RNA to RNA
#> Setting assay used for FindVariableFeatures.RNA to RNA
#> Setting assay used for ScaleData.RNA to RNA
#> Setting assay used for RunPCA.RNA to RNA
#> Setting assay used for JackStraw.RNA.pca to RNA
#> No assay information could be found for ScoreJackStraw
#> Warning: Adding a command log without an assay associated with it
#> Setting assay used for FindNeighbors.RNA.pca to RNA
#> No assay information could be found for FindClusters
#> Warning: Adding a command log without an assay associated with it
#> Setting assay used for RunUMAP.RNA.pca to RNA
#> Validating object structure for Assay5 'RNA'
#> Validating object structure for Graph 'RNA_nn'
#> Validating object structure for Graph 'RNA_snn'
#> Validating object structure for DimReduc 'pca'
#> Validating object structure for DimReduc 'umap'
#> Object representation is consistent with the most current Seurat version
```

```
pbmc3k.final
#> An object of class Seurat
#> 13714 features across 2638 samples within 1 assay
#> Active assay: RNA (13714 features, 2000 variable features)
#>  3 layers present: data, counts, scale.data
#>  2 dimensional reductions calculated: pca, umap
```

```
table(pbmc3k.final$seurat_annotations)
#>
#>  Naive CD4 T Memory CD4 T   CD14+ Mono            B        CD8 T FCGR3A+ Mono
#>          697          483          480          344          271          162
#>           NK           DC     Platelet
#>          155           32           14
```

```
pbmc3k.T <- subset(pbmc3k.final, subset = seurat_annotations %in% c("Naive CD4 T", "Memory CD4 T", "CD8
pbmc3k.T
#> An object of class Seurat
#> 13714 features across 1451 samples within 1 assay
#> Active assay: RNA (13714 features, 2000 variable features)
#>  3 layers present: data, counts, scale.data
#>  2 dimensional reductions calculated: pca, umap
```

Estimate marginal parameters

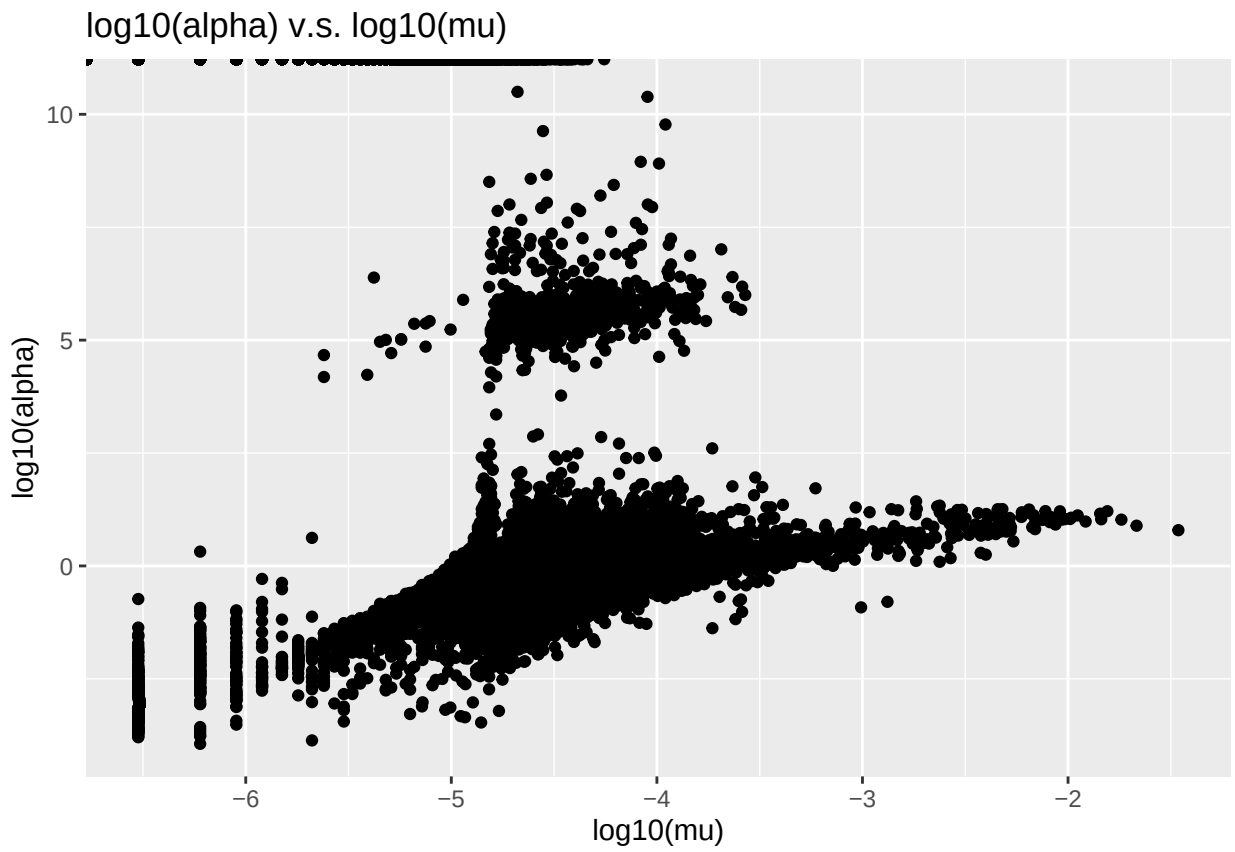
```
pbmc3k_count <- as.matrix(GetAssayData(object = pbmc3k.T, slot = "counts"))
#> Warning: The `slot` argument of `GetAssayData()` is deprecated as of SeuratObject 5.0.0.
#> i Please use the `layer` argument instead.
```

```
#> This warning is displayed once every 8 hours.
#> Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
#> generated.
```

```
marginal_para <- marginal_fit(pbmc3k_count)
#>
#> Attaching package: 'glmGamPoi'
#> The following object is masked from 'package:ggplot2':
#>
#>     vars
#> Make initial dispersion estimate
#> Make initial beta estimate
#> Estimate beta
#> Estimate dispersion
#> Fit dispersion trend
#> Shrink dispersion estimates
#> Estimate beta again
```

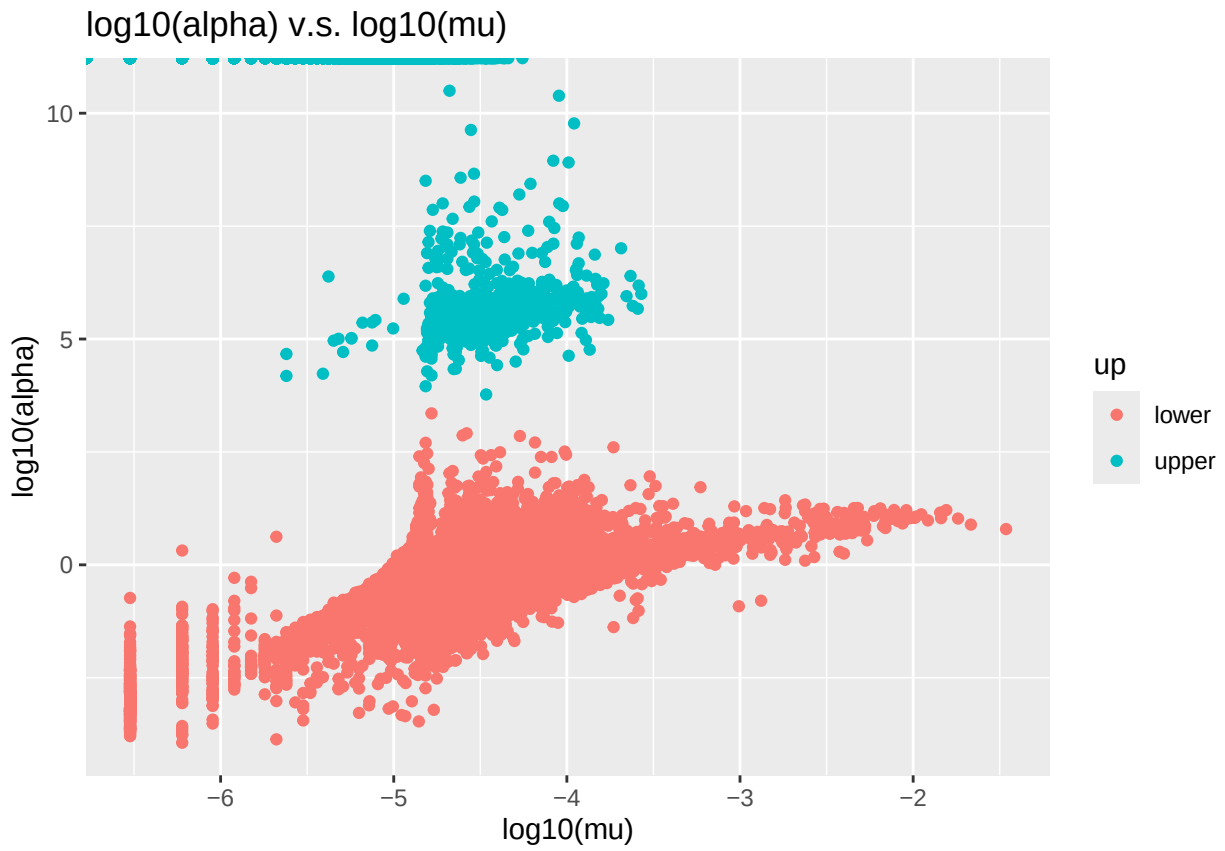
Check the relationship between mu and alpha.

```
ggplot(marginal_para, aes(x=mu, y=alpha))+
  geom_point()+labs(title = "log10(alpha) v.s. log10(mu)")+
  xlab("log10(mu)")+ylab("log10(alpha)")
```



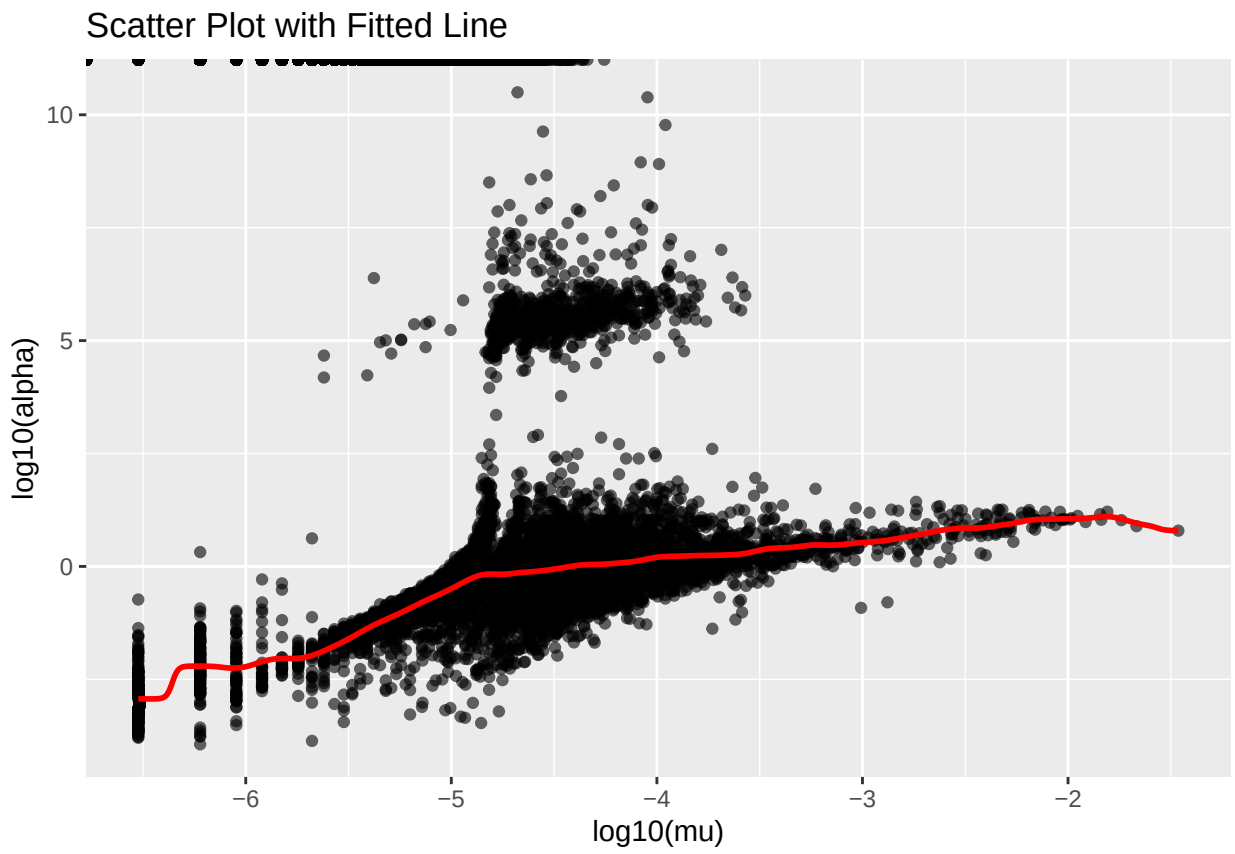
Fit a smooth curve between mu and alpha using the major cluster

```
marginal_para$up <- ifelse(marginal_para$alpha>3.5, "upper","lower")
ggplot(marginal_para, aes(x=mu, y=alpha, color=up))+
  geom_point()+labs(title = "log10(alpha) v.s. log10(mu)")+
  xlab("log10(mu)")+ylab("log10(alpha)")
```




```
## fit line (only use the major cluster)
marginal_para_sel <- marginal_para[marginal_para$up=="lower",]
## kernel smooth
km5 <- ksmooth(marginal_para_sel$mu, marginal_para_sel$alpha,
               kernel="normal", bandwidth = bw.SJ(marginal_para_sel$mu)*5)

ksmooth_df <- data.frame(mu = km5$x, alpha = km5$y)
# Plot the data and the fitted line
ggplot() +
  geom_point(data = marginal_para, aes(x = mu, y = alpha), alpha = 0.6) +
  geom_line(data = ksmooth_df, aes(x = mu, y = alpha), color = "red", size = 1) +
  labs(title = "Scatter Plot with Fitted Line",
       x = "log10(mu)",
       y = "log10(alpha)")
#> Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
#> i Please use `linewidth` instead.
#> This warning is displayed once every 8 hours.
#> Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
#> generated.
```



Use the fitted curve to find alpha.

```
log10mu <- marginal_para$mu
gene_name <- marginal_para$gene
names(log10mu) <- gene_name
mu <- 10^log10mu

fitted_trend <- data.frame(mu=km5$x, alpha=km5$y)
log10alpha <- rep(NA,nrow(marginal_para))
names(log10alpha) <- marginal_para$gene
for (i in 1:nrow(marginal_para)){
  idx <- which.min(abs(log10mu[i]-fitted_trend$mu))
  log10alpha[i] <- fitted_trend$alpha[idx]
}
alpha <- 10^log10alpha
```

Simulate independent data

```
simu_ind <- scSimu(mu, alpha, pbmc3k_count)
```

Simulate correlated data

Consider the correlation structure for highly expressed genes.

```
marginal_para <- marginal_para[order(marginal_para$mu, decreasing = T),]
cor_gene <- marginal_para$gene[1:500]
simu_cor <- scSimu(mu, alpha, pbmc3k_count, IND = F, cor_gene = cor_gene)
#> Warning: Data is of class matrix. Coercing to dgCMatrix.
#> [1] "IRLS converged after 4 iterations."
#> [1] "5 among 500 genes have negative variance estimates. Their co-expressions with other genes were
#> [1] "0.2846% co-expression estimates were greater than 1 and were set to 1."
#> [1] "0.2389% co-expression estimates were smaller than -1 and were set to -1."
#> <simpleError in chol.default(cor_mat): the leading minor of order 45 is not positive>
#> [1] "The correlation matrix is not positive semi-definite. Use eigen decomposition."
#> Warning in mvtnorm::rmvnorm(ncell, mean = rep(0, dim(cor_mat)[1]), sigma =
#> cor_mat, : sigma is numerically not positive semidefinite
```