countprop

Introduction

The `countprop` package allows estimation of several types of proportionality metrics for count-basedcompositional data such as 16S, metagenomic, and single-cell sequencing data. The package includes functions that allow standard empirical estimates of these proportionality metrics, as well as estimates based on the multinomial logit-normal model.

First, we'll define the model. Assume n samples and $J+1$ features. Suppose the counts for sample i for feature j are denoted by y_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, J+1$. They are modelled using the multinomial distribution:

$$y_i \sim \text{Multinomial}(M_i; p_{i1}, \ldots, p_{i(J+1)}),$$

where $M_i = \sum_{j=1}^{J+1} y_{ij}$ and proportion vector $\mathbf{p}_i = (p_{i1}, \ldots, p_{i(J+1)})$. The proportions themselves are modelled using a logit-normal model, which can be formulated through a set of latent vectors $(w_{i1}, \ldots, w_{iJ})$ which are related to the proportions by:

$$\begin{aligned} p_{ij} &= \text{alr}^{-1}(\mathbf{w}_i) \\ &= \begin{cases} \frac{\exp\{w_{ij}\}}{1+\sum_{j=1}^{J}\exp\{w_{ij}\}} & \text{if } j = 1, \ldots, J \\ \frac{1}{1+\sum_{j=1}^{J}\exp\{w_{ij}\}} & \text{if } j = J+1. \end{cases} \end{aligned}$$

The latent vectors are distributed as multivariate normal:

$$(w_{i1}, \ldots, w_{iJ}) \sim \text{MV-Normal}_J(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

The read-depths are assumed to be distributed as log-normal:

$$M_i \sim \text{Log-Normal}\left(\mu_\ell, \sigma_\ell^2\right).$$

Finally, to guard against spurious correlations, we apply the L_1-penalty to the inverse covariance matrix $\boldsymbol{\Sigma}$ (i.e. the "graphical lasso" penalty).

$$\ell(w_{i1}, \ldots, w_{iJ}) = \log\det\boldsymbol{\Sigma}^{-1} - \text{tr}(S\boldsymbol{\Sigma}^{-1}) - \lambda||\boldsymbol{\Sigma}^{-1}||_1$$

Fitting the model

The `countprop` package has a built-in function to estimate the model parameters. First, let's load the countprop library and look at the first few lines of the murine single cell sequencing dataset included with the package:

```
library(countprop)
#>
#> Attaching package: 'countprop'
#> The following object is masked from 'package:stats':
#>
#>     logLik
data(singlecell)

head(singlecell, 2)
#>                ENSMUSG00000064351 ENSMUSG00000064339 ENSMUSG00000064370
#> G1_cell1_count              40852              45108              31004
#> G1_cell2_count              67986              52596              57246
#>                ENSMUSG00000023944 ENSMUSG00000029580 ENSMUSG00000057113
#> G1_cell1_count              16235              19137              15962
#> G1_cell2_count              19273              20124              18578
#>                ENSMUSG00000037742 ENSMUSG00000020368 ENSMUSG00000064341
#> G1_cell1_count              11512               8614              17692
#> G1_cell2_count               9652              13785              24139
#>                ENSMUSG00000054766
#> G1_cell1_count               8902
#> G1_cell2_count              18429
```

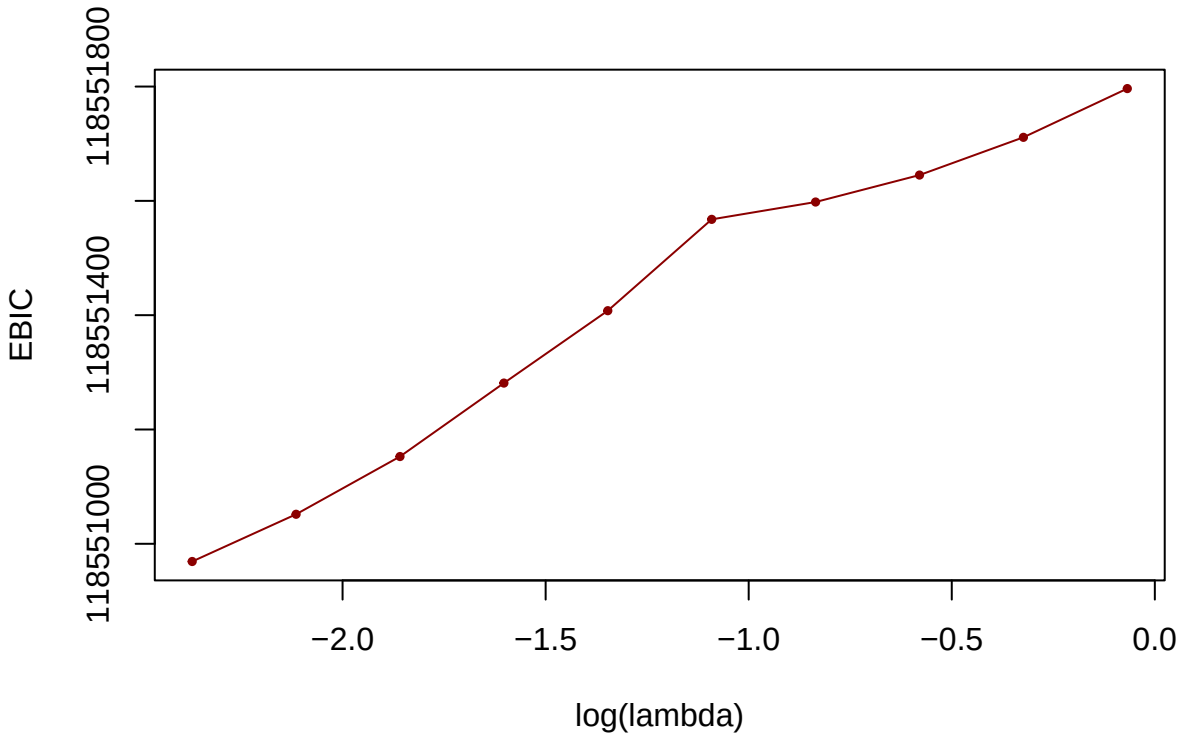
To fit the multinomial logit-normal model, we can use the `mleLR()` function:

```
mle <- mleLR(singlecell)

# Maximum likelihood estimates of model parameters
mle$mu
#> [1]  1.08972166  0.69105543  0.56031725  0.34323847  0.25345590  0.19768001
#> [7]  0.12912964 -0.02145714 -0.10901549
mle$Sigma.inv
#>               [,1]       [,2]       [,3]        [,4]      [,5]       [,6]
#>  [1,]  21.44710128 -9.1006784 -10.894588  -5.1174332  2.447743 -0.3932173
#>  [2,]  -9.10013470 19.1978605  -2.045736   1.7209571 -2.808991 -0.1172743
#>  [3,] -10.88951159 -2.0454178  27.490281   6.9191915  3.731340 -8.5116687
#>  [4,]  -5.11634377  1.7202705   6.919483  24.2221804 -2.833603 -3.9216127
#>  [5,]   2.44864368 -2.8074752   3.731312  -2.8334097 22.554437 -1.3602528
#>  [6,]  -0.39286959 -0.1168179  -8.513339  -3.9223575 -1.360460 20.6824940
#>  [7,]   2.90225574  0.2463541  -1.843747 -13.2077787 -3.662659 -4.6353533
#>  [8,]  -2.09514692 -0.6046498  -4.078285   0.2203073 -1.094891  2.4296618
#>  [9,]  -0.02713065 -5.8662579 -10.516366  -5.5981544 -5.101741  3.2801105
#>               [,7]       [,8]         [,9]
#>  [1,]   2.9031726 -2.0957806  -0.02558981
#>  [2,]   0.2477855 -0.6045823  -5.86419959
#>  [3,]  -1.8438336 -4.0771074 -10.51541009
#>  [4,] -13.2077869  0.2210259  -5.59699319
#>  [5,]  -3.6627851 -1.0945800  -5.10133297
#>  [6,]  -4.6355185  2.4295730   3.27986643
#>  [7,]  18.0886569 -1.0038532   4.16440953
```

```
#>  [8,]  -1.0034046  7.6764749   1.90561397
#>  [9,]   4.1645004  1.9059161  16.60315147
```

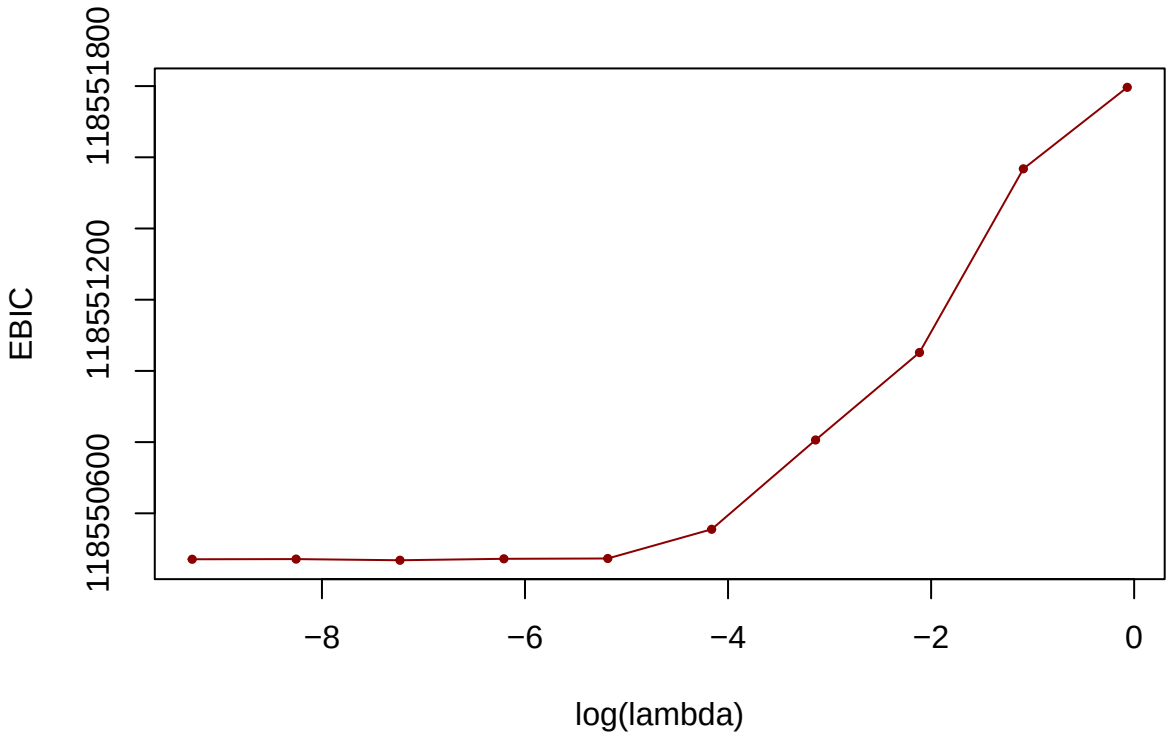
For the `mleLR()` function, it is necessary to specify a value for λ, which is the graphical lasso penalty parameter. The default is 0. However, we can also run multiple values of λ to find which one leads to the best fit based on the Extended Bayesian Information Criterion (EBIC). To do this, we use the `mlePath()` function. This allows us to choose the number of λ values we want to run the model on (`n.lambda` parameter). This can also be parallelized by setting `n.cores>1`. Once we've obtained the model fit, we can visualize the EBIC values for each λ value using `ebicPlot()`.

```
mle2 <- mlePath(singlecell, n.lambda=10, n.cores=1)
mle2$min.idx # Index of smallest lambda value
#> [1] 1

# Plot EBIC for different lambda values
ebicPlot(mle2)
```



In this case, the optimal value of λ is the one in the first position of the lambda vector. When the optimal λ value is the smallest one considered, then it's possible that an even smaller λ value would be optimal and was not considered. In this case, the argument `lambda.min.ratio` can be reduced from its default of `0.1`:

```
mle3 <- mlePath(singlecell, n.lambda=10, lambda.min.ratio = 0.0001, n.cores=1)
mle3$min.idx
#> [1] 3

ebicPlot(mle3)
```



The minimum EBIC now corresponds to the 3rd smallest value of λ.

Estimating the proportionality metrics

Once the model parameters have been estimated, the model-based proportionality metrics can be estimated:

```
# Variation matrix
logitNormalVariation(mle3$est.min$mu, mle3$est.min$Sigma)
#>             [,1]       [,2]       [,3]      [,4]       [,5]       [,6]
#>  [1,] 0.00000000 0.08223963 0.07155602 0.5100901 0.44032828 0.37281434
#>  [2,] 0.08223963 0.00000000 0.09422343 0.4724987 0.37644140 0.34607945
#>  [3,] 0.07155602 0.09422343 0.00000000 0.5050141 0.41572056 0.33307552
#>  [4,] 0.51009009 0.47249872 0.50501409 0.0000000 0.10530745 0.11702161
#>  [5,] 0.44032828 0.37644140 0.41572056 0.1053074 0.00000000 0.11564416
#>  [6,] 0.37281434 0.34607945 0.33307552 0.1170216 0.11564416 0.00000000
#>  [7,] 0.72512140 0.67375771 0.69687232 0.0886797 0.16033935 0.16848805
#>  [8,] 0.23966764 0.24571246 0.22236016 0.3764686 0.30275075 0.29647659
#>  [9,] 0.12782306 0.10277001 0.10009460 0.4856819 0.39130452 0.38111717
#> [10,] 0.41911591 0.37189939 0.39041370 0.1068904 0.06366782 0.08565271
#>            [,7]      [,8]      [,9]      [,10]
#>  [1,] 0.7251214 0.2396676 0.1278231 0.41911591
#>  [2,] 0.6737577 0.2457125 0.1027700 0.37189939
#>  [3,] 0.6968723 0.2223602 0.1000946 0.39041370
#>  [4,] 0.0886797 0.3764686 0.4856819 0.10689043
#>  [5,] 0.1603394 0.3027507 0.3913045 0.06366782
#>  [6,] 0.1684881 0.2964766 0.3811172 0.08565271
#>  [7,] 0.0000000 0.4978141 0.7138370 0.15792003
#>  [8,] 0.4978141 0.0000000 0.3012230 0.26633669
#>  [9,] 0.7138370 0.3012230 0.0000000 0.40606973
#> [10,] 0.1579200 0.2663367 0.4060697 0.00000000
```

```
# Phi matrix
logitNormalVariation(mle3$est.min$mu, mle3$est.min$Sigma, type="phi")
#>            [,1]      [,2]      [,3]      [,4]      [,5]      [,6]     [,7]
#>  [1,] 0.0000000 0.7122408 0.6197149 4.4176635 3.8134875 3.2287793 6.279954
#>  [2,] 0.7918107 0.0000000 0.9071919 4.5492611 3.6244124 3.3320848 6.487001
#>  [3,] 0.6498981 0.8557719 0.0000000 4.5867236 3.7757269 3.0251143 6.329251
#>  [4,] 2.7653214 2.5615295 2.7378032 0.0000000 0.5708971 0.6344023 0.480754
#>  [5,] 3.1630499 2.7041255 2.9862830 0.7564645 0.0000000 0.8307171 1.151780
#>  [6,] 3.2627902 3.0288123 2.9150047 1.0241477 1.0120926 0.0000000 1.474571
#>  [7,] 2.3036213 2.1404452 2.2138775 0.2817245 0.5093784 0.5352658 0.000000
#>  [8,] 1.7165531 1.7598474 1.5925930 2.6963520 2.1683684 2.1234315 3.565456
#>  [9,] 0.9386991 0.7547160 0.7350685 3.5667203 2.8736378 2.7988245 5.242232
#> [10,] 3.2627271 2.8951566 3.0392866 0.8321190 0.4956403 0.6667879 1.229373
#>           [,8]      [,9]     [,10]
#>  [1,] 2.075655 1.1070187 3.6297766
#>  [2,] 2.365742 0.9894791 3.5806815
#>  [3,] 2.019557 0.9090959 3.5458807
#>  [4,] 2.040927 2.6329985 0.5794788
#>  [5,] 2.174777 2.8108931 0.4573508
#>  [6,] 2.594699 3.3354547 0.7496139
#>  [7,] 1.581494 2.2677722 0.5016925
#>  [8,] 0.000000 2.1574264 1.9075628
#>  [9,] 2.212103 0.0000000 2.9820696
#> [10,] 2.073374 3.1611654 0.0000000

# Rho matrix
logitNormalVariation(mle3$est.min$mu, mle3$est.min$Sigma, type="rho")
#>              [,1]         [,2]       [,3]       [,4]        [,5]       [,6]
#>  [1,]  1.00000000  0.625039507  0.6827761 -0.7007218 -0.72897392 -0.6228478
#>  [2,]  0.62503951  1.000000000  0.5596340 -0.6387863 -0.54867777 -0.5866094
#>  [3,]  0.68277613  0.559634000  1.0000000 -0.7144516 -0.66746118 -0.4845195
#>  [4,] -0.70072184 -0.638786336 -0.7144516  1.0000000  0.67464527  0.6082592
#>  [5,] -0.72897392 -0.548677770 -0.6674612  0.6746453  1.00000000  0.5437605
#>  [6,] -0.62284781 -0.586609446 -0.4845195  0.6082592  0.54376047  1.0000000
#>  [7,] -0.68538580 -0.609406742 -0.6401703  0.8223685  0.64681764  0.6072878
#>  [8,]  0.06044927 -0.009151578  0.1095796 -0.1616495 -0.08578393 -0.1677645
#>  [9,]  0.49203285  0.571850804  0.5935639 -0.5147734 -0.42095958 -0.5218336
#> [10,] -0.71823925 -0.600817310 -0.6365488  0.6584046  0.76213682  0.6471104
#>              [,7]         [,8]        [,9]        [,10]
#>  [1,] -0.68538580  0.060449268  0.49203285 -0.718239252
#>  [2,] -0.60940674 -0.009151578  0.57185080 -0.600817310
#>  [3,] -0.64017033  0.109579568  0.59356390 -0.636548798
#>  [4,]  0.82236854 -0.161649476 -0.51477341  0.658404599
#>  [5,]  0.64681764 -0.085783932 -0.42095958  0.762136817
#>  [6,]  0.60728782 -0.167764509 -0.52183365  0.647110396
#>  [7,]  1.00000000 -0.095551268 -0.58298022  0.643706573
#>  [8,] -0.09555127  1.000000000 -0.09221128  0.006492412
#>  [9,] -0.58298022 -0.092211282  1.00000000 -0.534503448
#> [10,]  0.64370657  0.006492412 -0.53450345  1.000000000
```

The package also provides the standard naive (empirical) estimates of the proportionality metrics.

```
# Naive (empirical) variation matrix
naiveVariation(singlecell)
#>                     ENSMUSG00000064351 ENSMUSG00000064339 ENSMUSG00000064370
#> ENSMUSG00000064351          0.00000000          0.08152267          0.06974364
#> ENSMUSG00000064339          0.08152267          0.00000000          0.09152832
#> ENSMUSG00000064370          0.06974364          0.09152832          0.00000000
#> ENSMUSG00000023944          0.50872551          0.47281649          0.50513276
#> ENSMUSG00000029580          0.44476070          0.37417054          0.41682223
#> ENSMUSG00000057113          0.37502706          0.35788732          0.34008017
#> ENSMUSG00000037742          0.72592022          0.67517400          0.69456335
#> ENSMUSG00000020368          0.23904178          0.25093706          0.22512118
#> ENSMUSG00000064341          0.12486115          0.10187624          0.09768942
#> ENSMUSG00000054766          0.41902483          0.37227665          0.38995197
#>                     ENSMUSG00000023944 ENSMUSG00000029580 ENSMUSG00000057113
#> ENSMUSG00000064351          0.5087255          0.44476070          0.3750271
#> ENSMUSG00000064339          0.4728165          0.37417054          0.3578873
#> ENSMUSG00000064370          0.5051328          0.41682223          0.3400802
#> ENSMUSG00000023944          0.0000000          0.10347647          0.1255767
#> ENSMUSG00000029580          0.1034765          0.00000000          0.1316500
#> ENSMUSG00000057113          0.1255767          0.13165003          0.0000000
#> ENSMUSG00000037742          0.0865833          0.16039549          0.1731814
#> ENSMUSG00000020368          0.3830622          0.31195108          0.2972780
#> ENSMUSG00000064341          0.4840748          0.39262793          0.3869858
#> ENSMUSG00000054766          0.1068193          0.06553481          0.0929203
#>                     ENSMUSG00000037742 ENSMUSG00000020368 ENSMUSG00000064341
#> ENSMUSG00000064351          0.7259202          0.2390418          0.12486115
#> ENSMUSG00000064339          0.6751740          0.2509371          0.10187624
#> ENSMUSG00000064370          0.6945633          0.2251212          0.09768942
#> ENSMUSG00000023944          0.0865833          0.3830622          0.48407480
#> ENSMUSG00000029580          0.1603955          0.3119511          0.39262793
#> ENSMUSG00000057113          0.1731814          0.2972780          0.38698581
#> ENSMUSG00000037742          0.0000000          0.4988237          0.71450463
#> ENSMUSG00000020368          0.4988237          0.0000000          0.30353326
#> ENSMUSG00000064341          0.7145046          0.3035333          0.00000000
#> ENSMUSG00000054766          0.1576063          0.2696139          0.40586212
#>                     ENSMUSG00000054766
#> ENSMUSG00000064351          0.41902483
#> ENSMUSG00000064339          0.37227665
#> ENSMUSG00000064370          0.38995197
#> ENSMUSG00000023944          0.10681931
#> ENSMUSG00000029580          0.06553481
#> ENSMUSG00000057113          0.09292030
#> ENSMUSG00000037742          0.15760629
#> ENSMUSG00000020368          0.26961390
#> ENSMUSG00000064341          0.40586212
#> ENSMUSG00000054766          0.00000000

# Naive (empirical) Phi matrix
naiveVariation(singlecell, type="phi")
#>                     ENSMUSG00000064351 ENSMUSG00000064339 ENSMUSG00000064370
#> ENSMUSG00000064351          0.0000000          0.6293701          0.5384338
#> ENSMUSG00000064339          0.7010116          0.0000000          0.7870500
#> ENSMUSG00000064370          0.5798331          0.7609461          0.0000000
```

```
#> ENSMUSG00000023944          3.0723239          2.8554602          3.0506264
#> ENSMUSG00000029580          3.6231064          3.0480653          3.3955142
#> ENSMUSG00000057113          3.4668646          3.3084196          3.1438049
#> ENSMUSG00000037742          2.4491384          2.2779288          2.3433454
#> ENSMUSG00000020368          1.8244629          1.9152524          1.7182153
#> ENSMUSG00000064341          0.8603142          0.7019443          0.6730964
#> ENSMUSG00000054766          3.6941337          3.2820005          3.4378266
#>                     ENSMUSG00000023944 ENSMUSG00000029580 ENSMUSG00000057113
#> ENSMUSG00000064351          3.9274550          3.4336349          2.8952782
#> ENSMUSG00000064339          4.0657383          3.2174840          3.0774650
#> ENSMUSG00000064370          4.1995612          3.4653671          2.8273507
#> ENSMUSG00000023944          0.0000000          0.6249210          0.7583901
#> ENSMUSG00000029580          0.8429393          0.0000000          1.0724465
#> ENSMUSG00000057113          1.1608697          1.2170131          0.0000000
#> ENSMUSG00000037742          0.2921182          0.5411486          0.5842864
#> ENSMUSG00000020368          2.9236844          2.3809360          2.2689454
#> ENSMUSG00000064341          3.3353562          2.7052720          2.6663968
#> ENSMUSG00000054766          0.9417218          0.5777566          0.8191877
#>                     ENSMUSG00000037742 ENSMUSG00000020368 ENSMUSG00000064341
#> ENSMUSG00000064351          5.6042384           1.845447          0.9639512
#> ENSMUSG00000064339          5.8058060           2.157802          0.8760315
#> ENSMUSG00000064370          5.7744449           1.871607          0.8121680
#> ENSMUSG00000023944          0.5228988           2.313411          2.9234520
#> ENSMUSG00000029580          1.3066125           2.541214          3.1984227
#> ENSMUSG00000057113          1.6009421           2.748129          3.5774150
#> ENSMUSG00000037742          0.0000000           1.682951          2.4106240
#> ENSMUSG00000020368          3.8072229           0.000000          2.3166878
#> ENSMUSG00000064341          4.9230562           2.091395          0.0000000
#> ENSMUSG00000054766          1.3894611           2.376923          3.5780909
#>                     ENSMUSG00000054766
#> ENSMUSG00000064351          3.2349492
#> ENSMUSG00000064339          3.2011985
#> ENSMUSG00000064370          3.2419738
#> ENSMUSG00000023944          0.6451092
#> ENSMUSG00000029580          0.5338592
#> ENSMUSG00000057113          0.8589836
#> ENSMUSG00000037742          0.5317383
#> ENSMUSG00000020368          2.0578016
#> ENSMUSG00000064341          2.7964578
#> ENSMUSG00000054766          0.0000000

# Naive (empirical) Rho matrix
naiveVariation(singlecell, type="rho")
#>                     ENSMUSG00000064351 ENSMUSG00000064339 ENSMUSG00000064370
#> ENSMUSG00000064351          1.00000000          0.66836904          0.7208164
#> ENSMUSG00000064339          0.66836904          1.00000000          0.6131110
#> ENSMUSG00000064370          0.72081642          0.61311104          1.0000000
#> ENSMUSG00000023944         -0.72382787         -0.67739068         -0.7670291
#> ENSMUSG00000029580         -0.76291350         -0.56524206         -0.7150425
#> ENSMUSG00000057113         -0.57769763         -0.59438295         -0.4885961
#> ENSMUSG00000037742         -0.70432302         -0.63602752         -0.6668968
#> ENSMUSG00000020368          0.08255256         -0.01465269          0.1041830
#> ENSMUSG00000064341          0.54540557          0.61030751          0.6319394
```

```
#> ENSMUSG00000054766        -0.72466327        -0.62054800         -0.6685145
#>                    ENSMUSG00000023944 ENSMUSG00000029580 ENSMUSG00000057113
#> ENSMUSG00000064351         -0.7238279         -0.7629135         -0.5776976
#> ENSMUSG00000064339         -0.6773907         -0.5652421         -0.5943829
#> ENSMUSG00000064370         -0.7670291         -0.7150425         -0.4885961
#> ENSMUSG00000023944          1.0000000          0.6411304          0.5412856
#> ENSMUSG00000029580          0.6411304          1.0000000          0.4299172
#> ENSMUSG00000057113          0.5412856          0.4299172          1.0000000
#> ENSMUSG00000037742          0.8125828          0.6173360          0.5719401
#> ENSMUSG00000020368         -0.2914952         -0.2292326         -0.2428268
#> ENSMUSG00000064341         -0.5579250         -0.4656251         -0.5277219
#> ENSMUSG00000054766          0.6171530          0.7225294          0.5806931
#>                    ENSMUSG00000037742 ENSMUSG00000020368 ENSMUSG00000064341
#> ENSMUSG00000064351         -0.7043230         0.08255256          0.54540557
#> ENSMUSG00000064339         -0.6360275        -0.01465269          0.61030751
#> ENSMUSG00000064370         -0.6668968         0.10418296          0.63193936
#> ENSMUSG00000023944          0.8125828        -0.29149517         -0.55792499
#> ENSMUSG00000029580          0.6173360        -0.22923264         -0.46562514
#> ENSMUSG00000057113          0.5719401        -0.24282682         -0.52772188
#> ENSMUSG00000037742          1.0000000        -0.16706139         -0.61823768
#> ENSMUSG00000020368         -0.1670614         1.00000000         -0.09914205
#> ENSMUSG00000064341         -0.6182377        -0.09914205          1.00000000
#> ENSMUSG00000054766          0.6154331        -0.10294019         -0.56967664
#>                    ENSMUSG00000054766
#> ENSMUSG00000064351         -0.7246633
#> ENSMUSG00000064339         -0.6205480
#> ENSMUSG00000064370         -0.6685145
#> ENSMUSG00000023944          0.6171530
#> ENSMUSG00000029580          0.7225294
#> ENSMUSG00000057113          0.5806931
#> ENSMUSG00000037742          0.6154331
#> ENSMUSG00000020368         -0.1029402
#> ENSMUSG00000064341         -0.5696766
#> ENSMUSG00000054766          1.0000000
```