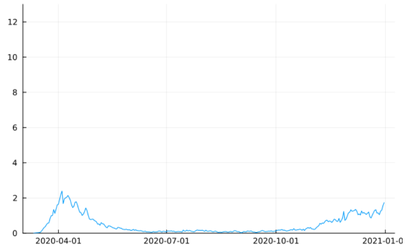
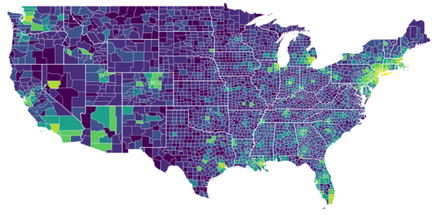
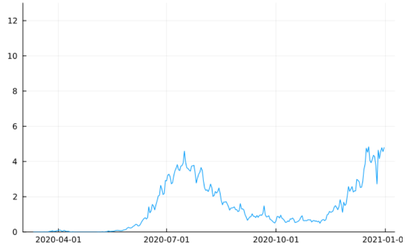
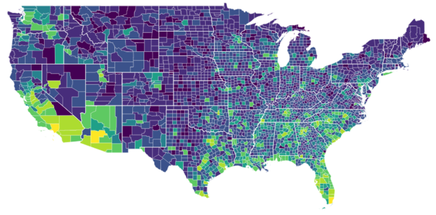
K=1


12
10
8
6
4
2
0
2020-04-01
2020-07-01
2020-10-01
2021-01-0

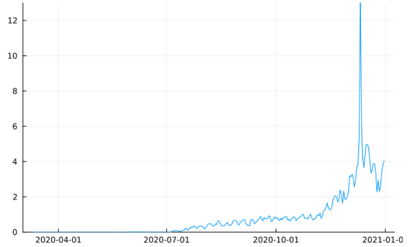

> 100
10 - 100
2 - 10
1 - 2
0.5 - 1
0.1 - 0.5
0.01 - 0.1
< 0.01

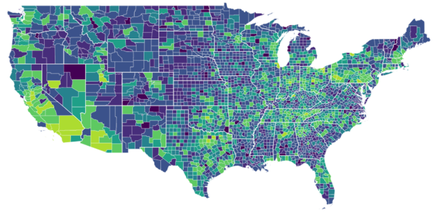
K=2


12
10
8
6
4
2
0
2020-04-01
2020-07-01
2020-10-01
2021-01-0

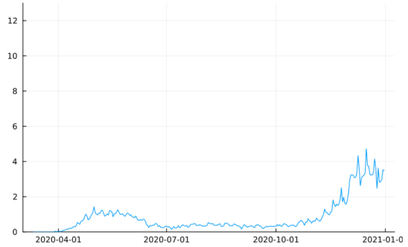

> 100
10 - 100
2 - 10
1 - 2
0.5 - 1
0.1 - 0.5
0.01 - 0.1
< 0.01

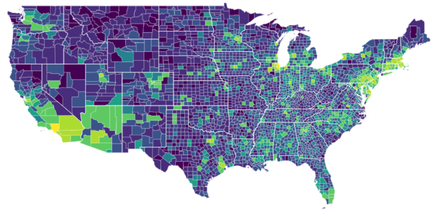
K=3


12
10
8
6
4
2
0
2020-04-01
2020-07-01
2020-10-01
2021-01-0

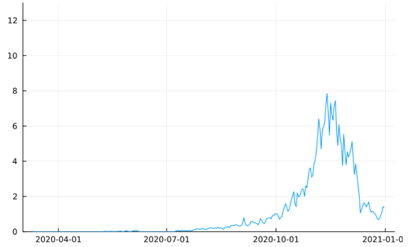

10 - 100
2 - 10
1 - 2
0.5 - 1
0.1 - 0.5
0.01 - 0.1
< 0.01

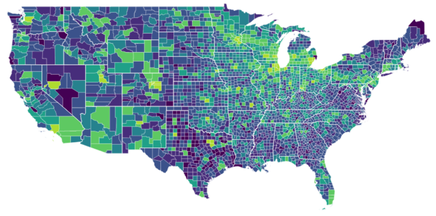
K=4


12
10
8
6
4
2
0
2020-04-01
2020-07-01
2020-10-01
2021-01-0

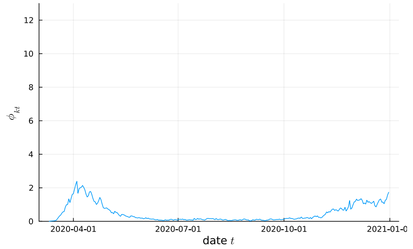

> 100
10 - 100
2 - 10
1 - 2
0.5 - 1
0.1 - 0.5
0.01 - 0.1
< 0.01

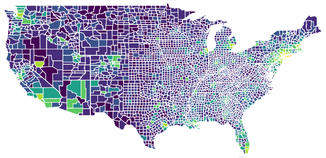
K=5


12
10
8
6
4
2
0
2020-04-01
2020-07-01
2020-10-01
2021-01-0

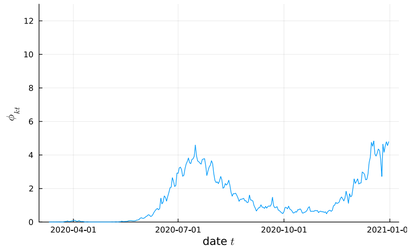

10 - 100
2 - 10
1 - 2
0.5 - 1
0.1 - 0.5
0.01 - 0.1
< 0.01

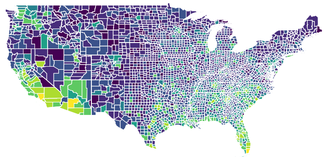
Latent Class K=1



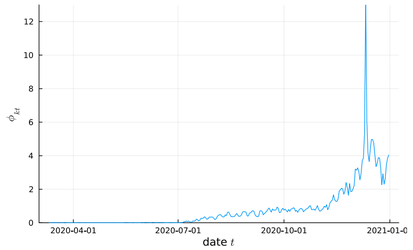


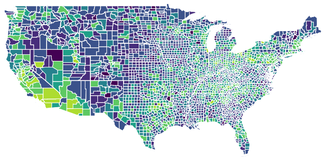
Latent Class K=2



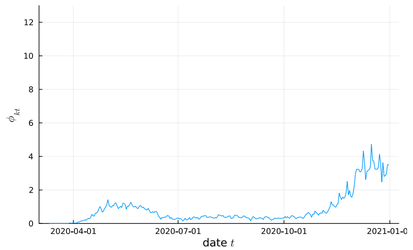


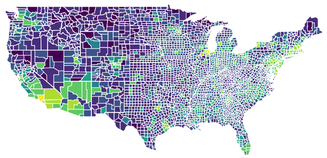
Latent Class K=3



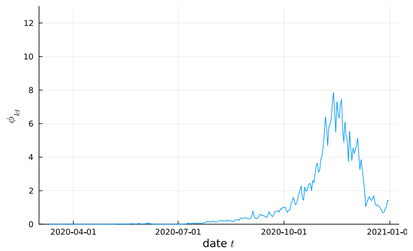


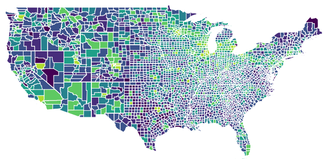
Latent Class K=4

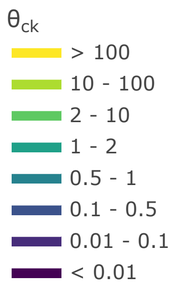


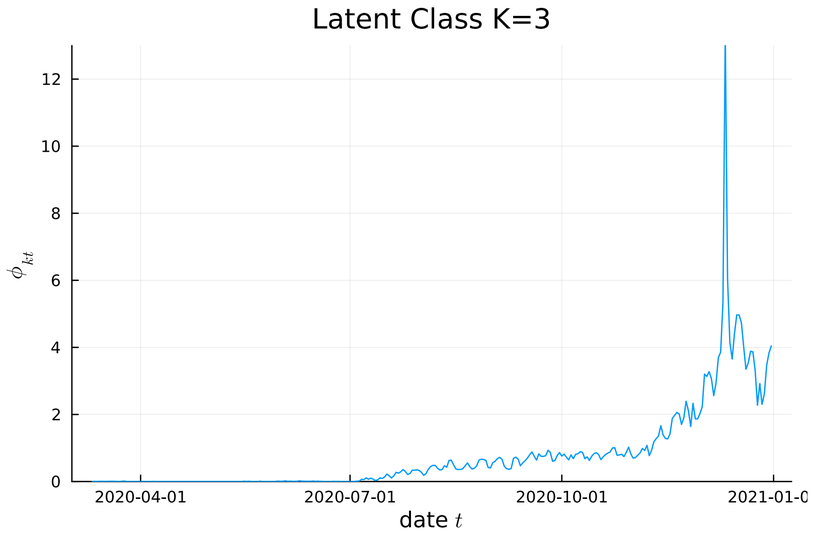


Latent Class K=5








Latent Class K=3
12
10
8
6
4
2
0
ϕ_{kt}
2020-04-01
2020-07-01
2020-10-01
2021-01-0
date t

Key takeaways

- An underdispersned likelihood achieves better holdout performance (on two tasks) than the standard equidispersed likelihood
- We showcase the median Poisson likelihood and justify why it is necessary (max forecasting results for appendix?)
- We motivate why one would want a model that captures latent structure in this setting
  - Is the structure we present good enough?

Example 3: RNA-sequencing data

Data

	Gene 1	Gene 2	Gene 3	…
Subject 1	0	223	324	…
Subject 2	1234	546	0	…
…	…	…	…	…

$$Y_{gs} \sim MaxNB\left(\sum_{k=1}^{K}\theta_{gk}\phi_{ks}, \sigma\left(\sum_{q=1}^{Q}\beta_{gq}\tau_{qs}\right), D\right)$$

$$\theta_{gk} \sim \Gamma(a,b), \phi_{kt} \sim \Gamma(c,d)$$

$$\beta_{gq} \sim Normal(0,1), \phi_{qt} \sim Normal(0,1)$$

Simulation Settings

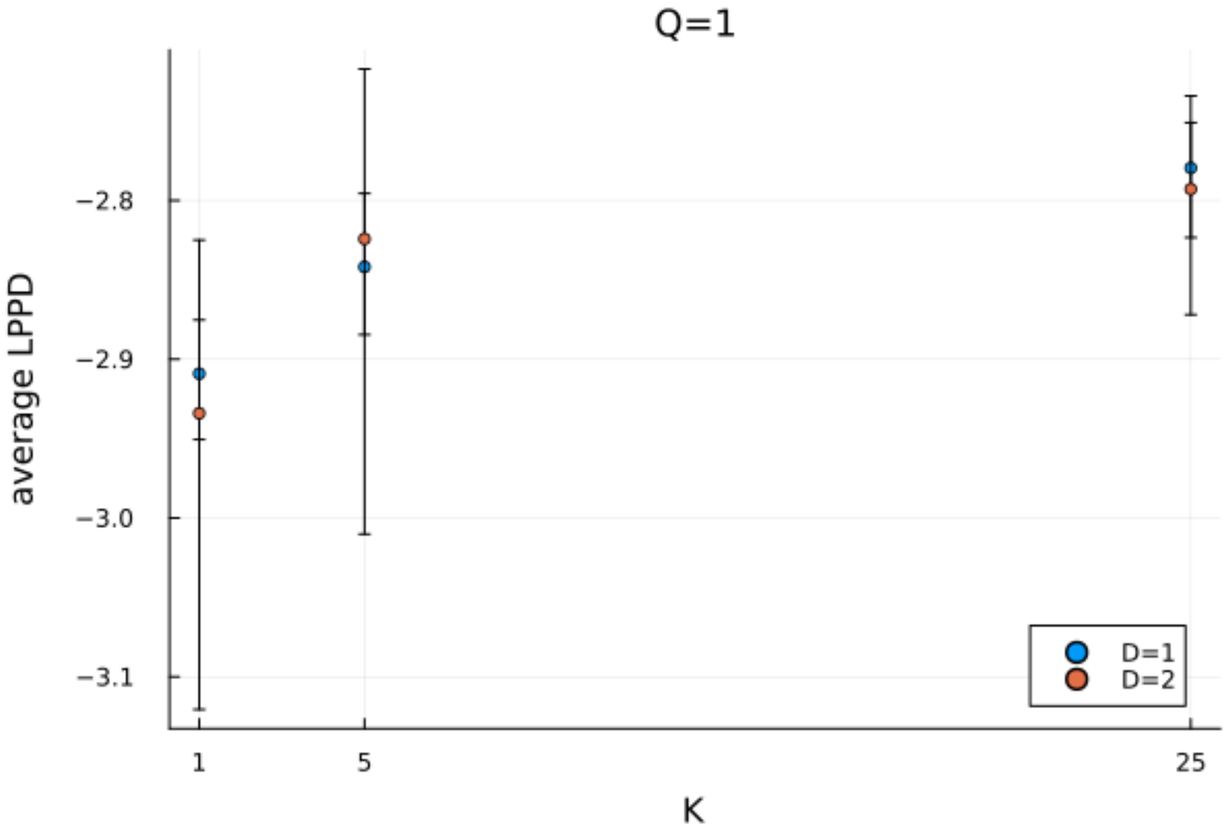
- We burned in for 10000 iterations and took 100 samples thinned by 10 iterations.
- We used 4 masks (should be 5), and ran only one chain each (the plan is for 4 chains)

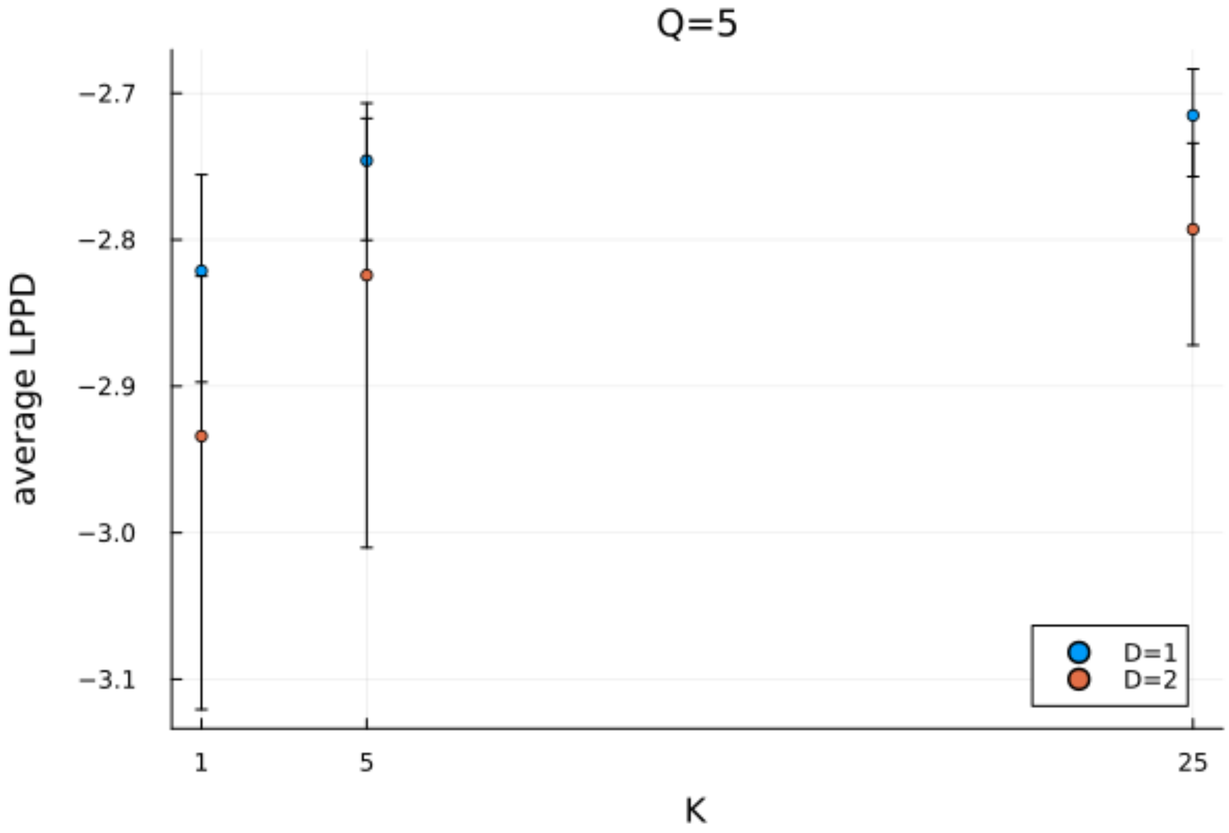
$$Y_{gs} \sim MaxNB\left(\sum_{k=1}^{K}\theta_{gk}\phi_{ks}, \sigma\left(\sum_{q=1}^{Q}\beta_{gq}\tau_{qs}\right), D\right)$$

$$\theta_{gk} \sim \Gamma(a,b), \phi_{kt} \sim \Gamma(c,d) \qquad \beta_{gq} \sim Normal(0,1), \phi_{qt} \sim Normal(0,1)$$

- We set a=b=c=d=1, $K \in [5,10,25]$, $Q \in [1,5]$, $D \in [1,2]$
- We holdout 2.5% of the entries in the matrix

LPPD results





To me, this suggest burnin is insufficient.

Question: is RNA-sequencing data conditionally underdispersed

- We will test this hypothesis as follows.
  - First run the model on 3000 random genes with nothing heldout.
  - We burnin 10000, only keep one sample (unfortunately). Only one run each setting.
  - We used all combinations of $K \in [5,10,25,50]$, $Q \in [1,5,10]$, $D \in [1,2]$.
- For every data-point,
  - calculate the fitted r and p
  - Simulate 5000 times from the likelihood $MaxNB(r, p, D)$ or $NB(r, p)$.
  - Calculate F[x] = V[x]/E[x]
  - Calculate proportion of times this quantity is less than 1.
  - Then threshold by the size of the counts Y to see if large counts are underdispersed

Proportion Underdispersed

Of all counts in the matrix

D=2

K	Q	F < 1	F < .95	F < .9
1.00000	**10.00000**	0.12925	0.00287	0.00049
1.00000	**1.00000**	0.06988	0.00006	0.00000
1.00000	**5.00000**	0.11121	0.00088	0.00005
25.00000	**10.00000**	**0.26681**	**0.01987**	**0.00465**
25.00000	**1.00000**	0.18403	0.00937	0.00297
25.00000	**5.00000**	0.25573	0.01605	0.00357
50.00000	**1.00000**	0.17734	0.01106	0.00362
50.00000	**5.00000**	**0.27130**	**0.02090**	**0.00519**
5.00000	**10.00000**	0.22008	0.01085	0.00226
5.00000	**5.00000**	0.19280	0.00450	0.00074

D=1

K	Q	all1	all95	all9
1.00000	**10.00000**	0.10781	0.00000	0.00000
1.00000	**1.00000**	0.06679	0.00000	0.00000
1.00000	**5.00000**	0.10896	0.00000	0.00000
25.00000	**10.00000**	0.17392	0.00001	0.00000
25.00000	**1.00000**	0.11668	0.00000	0.00000
25.00000	**5.00000**	0.16299	0.00000	0.00000
50.00000	**10.00000**	0.17124	0.00000	0.00000
50.00000	**1.00000**	0.09595	0.00000	0.00000
50.00000	**5.00000**	0.16566	0.00001	0.00000
5.00000	**10.00000**	0.15515	0.00000	0.00000
5.00000	**1.00000**	0.09366	0.00000	0.00000
5.00000	**5.00000**	0.14747	0.00000	0.00000

Proportion Underdispersed

Of only larger counts

D=2

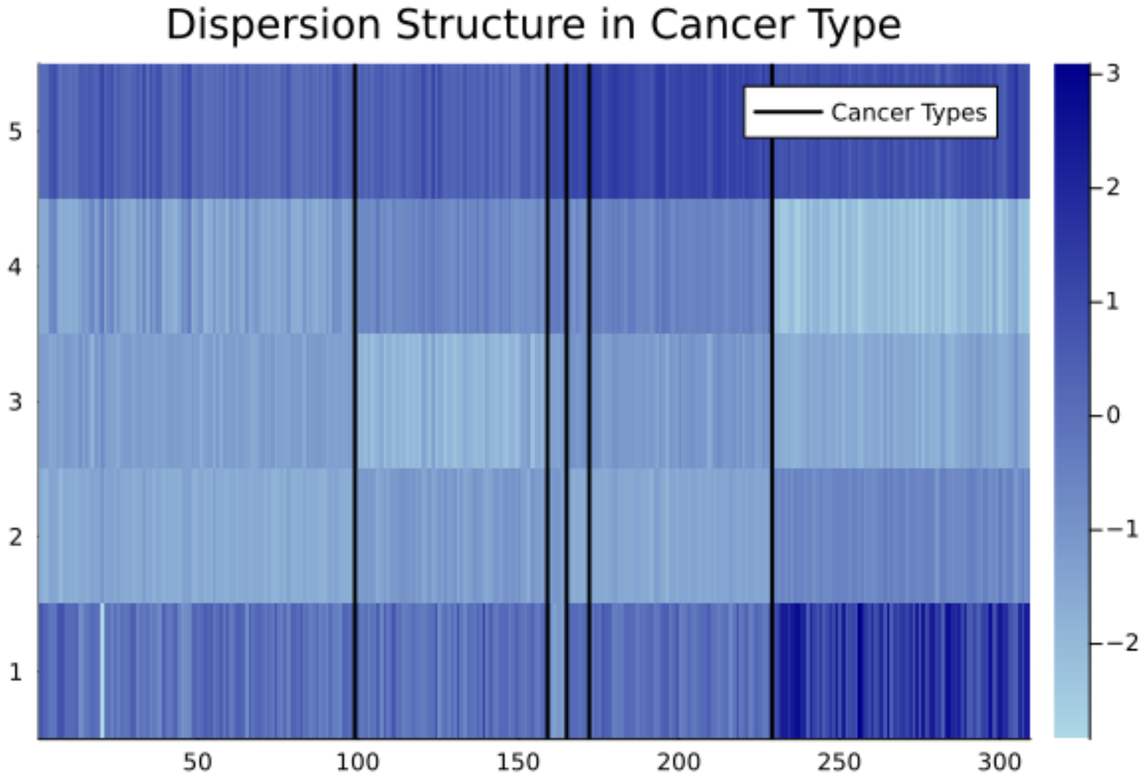
K	Q	five	ten	fifty
1.00000	**10.00000**	0.00003	0.00000	0.00000
1.00000	**1.00000**	0.00000	0.00000	0.00000
1.00000	**5.00000**	0.00001	0.00000	0.00000
25.00000	**10.00000**	0.00030	0.00001	0.00000
25.00000	**1.00000**	**0.00052**	**0.00013**	**0.00001**
25.00000	**5.00000**	0.00026	0.00001	0.00000
50.00000	**1.00000**	**0.00063**	**0.00017**	**0.00001**
50.00000	**5.00000**	0.00041	0.00003	0.00000
5.00000	**10.00000**	0.00011	0.00001	0.00000
5.00000	**5.00000**	0.00005	0.00000	0.00000

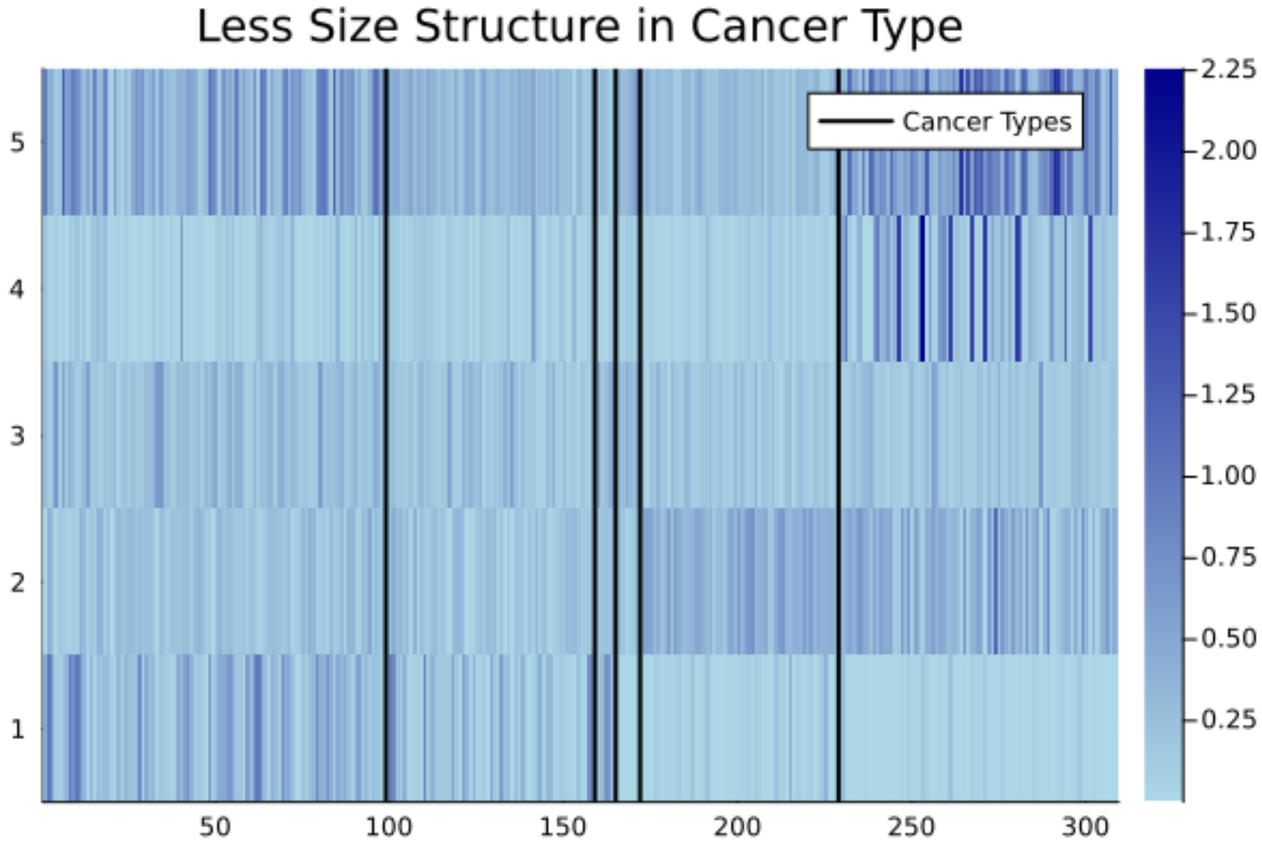
D=1

K	Q	five	ten	fifty
1.00000	**10.00000**	0.00000	0.00000	0.00000
1.00000	**1.00000**	0.00000	0.00000	0.00000
1.00000	**5.00000**	0.00000	0.00000	0.00000
25.00000	**10.00000**	0.00000	0.00000	0.00000
25.00000	**1.00000**	0.00000	0.00000	0.00000
25.00000	**5.00000**	0.00000	0.00000	0.00000
50.00000	**10.00000**	0.00000	0.00000	0.00000
50.00000	**1.00000**	0.00000	0.00000	0.00000
50.00000	**5.00000**	0.00000	0.00000	0.00000
5.00000	**10.00000**	0.00000	0.00000	0.00000
5.00000	**1.00000**	0.00000	0.00000	0.00000
5.00000	**5.00000**	0.00000	0.00000	0.00000

Structure: Cancer Types

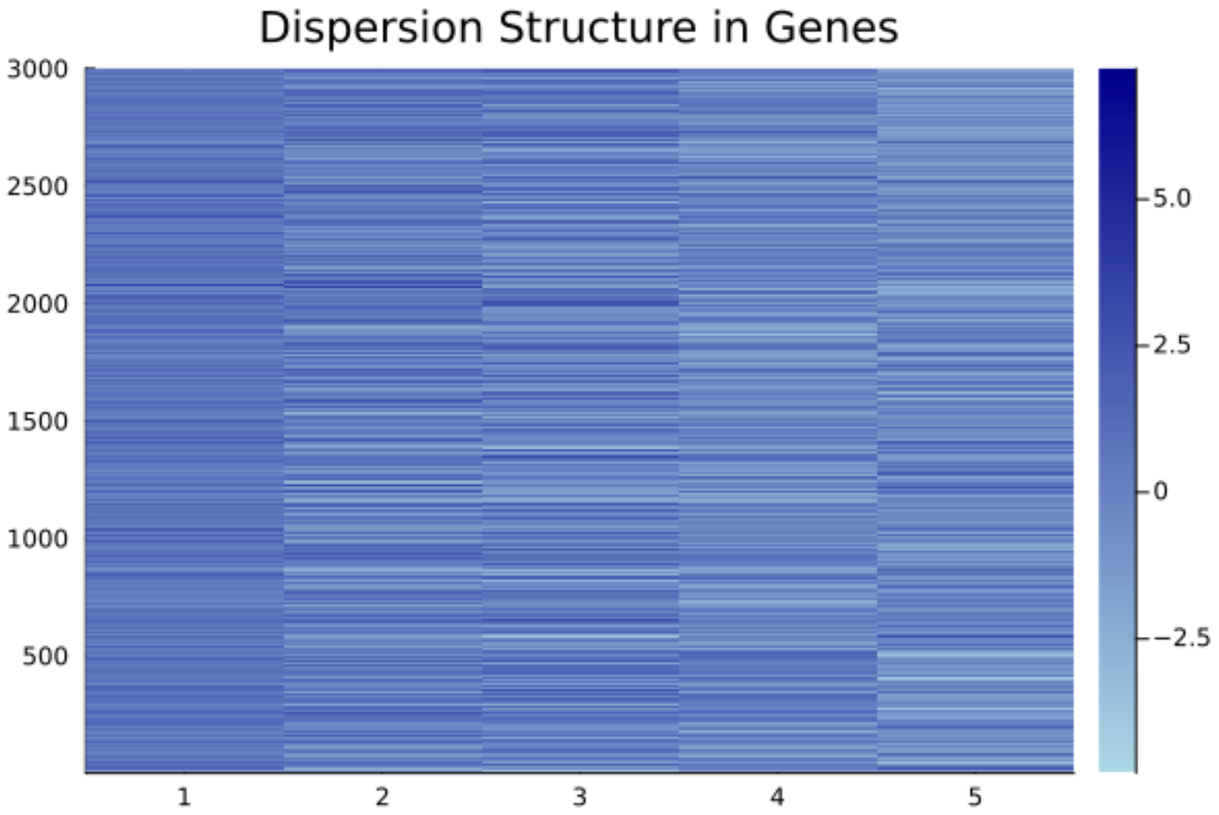
K=5,Q=5

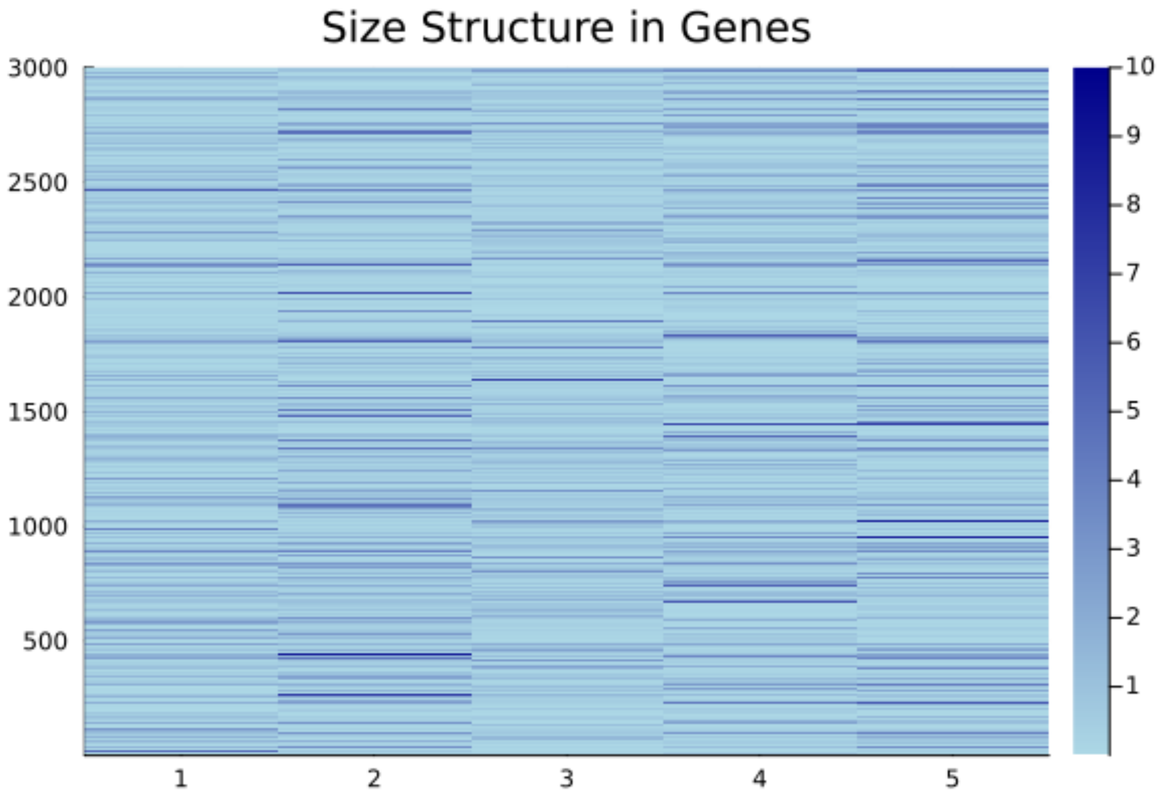




Structure: Genes

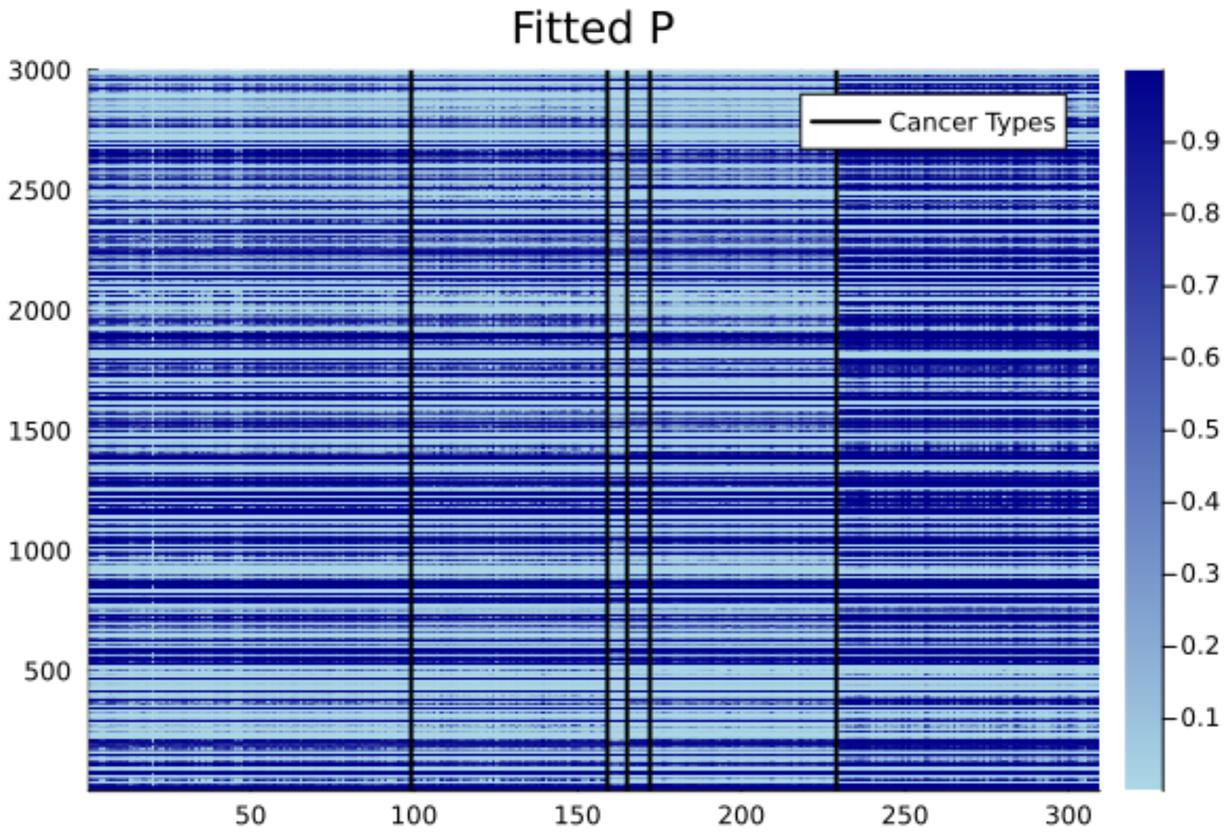
K=5,Q=5

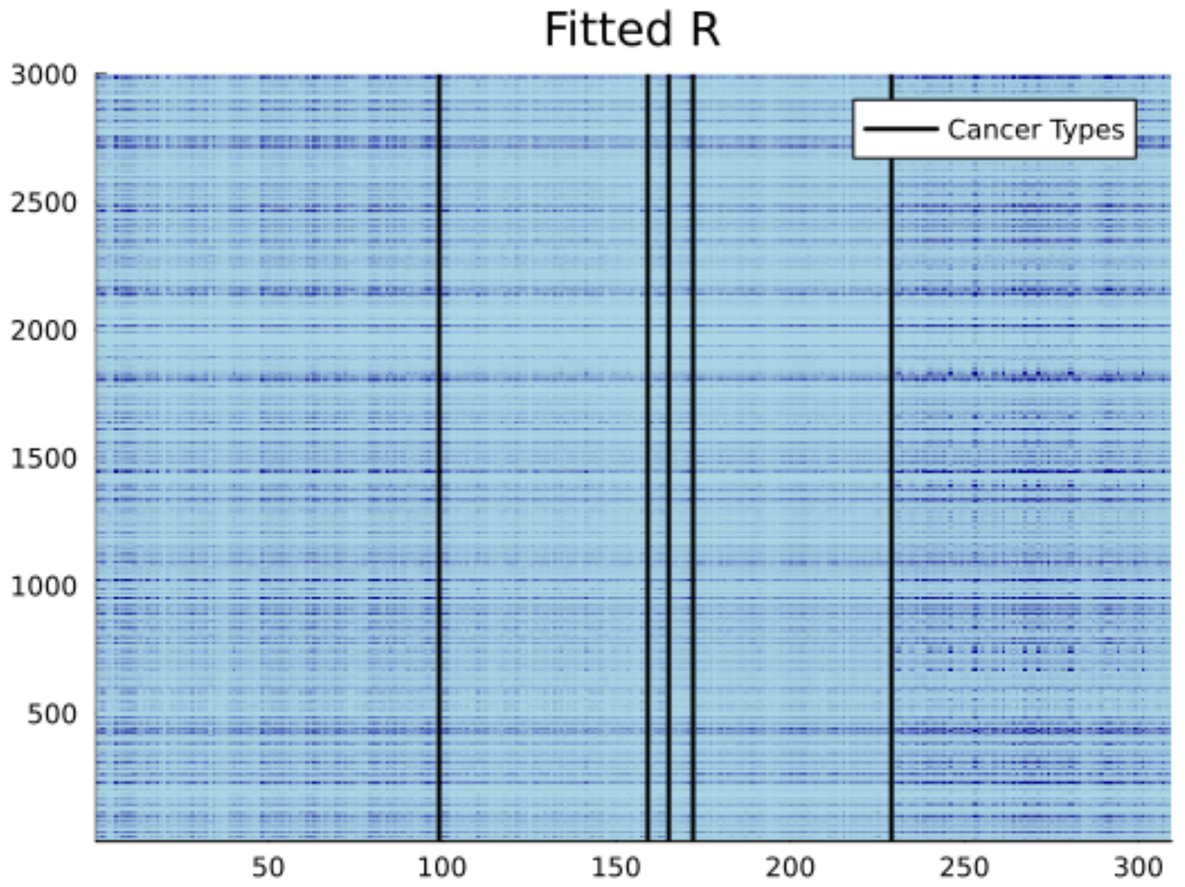




Structure: Fitted P and R

K=5,Q=5





Takeaways

- Limited evidence that RNA-sequencing data is underdispersed in our model class
    - However, there is always the possibility that we are not conditioning on the right structure to reveal the conditional underdispersion
- Regardless, the NB model class gives us the tools to test this sort of hypothesis.
- It also gives us the tools to model both over and underdispersion flexibly
- Factorizing the dispersion parameter yields observed differences among cancer types

Next

- Whatever we said we'd change as we went through this
- Run jobs
- Median issue
- Lay out synthetic experiments
- Writing
- Set an ambitious goal
- (Briefly discuss KB and OMD projects)
  - Text based ideal points