Week 4

Concern that problems are due to high dimensionality. Focus this week: Filter to one Genus/family/something and repeat.

Start thinking/writing about motivations for needing/wanting to model everything at once instead of running separately.

What dataset to use? Prevalence filter first? Going to use the 10p filtered dataset and then pick a genus.

Also looking into R ways to resolve phylogenetic tree to genus level. Potentially group clade? https://yulab-smu.top/treedata-book/chapter6.html

(Needed to add phylogenetic tree to saved rds.)

Meeting notes

Found bug. Ugh. Variance term was incorrect. Check things now.

Try independence, Try fixed values. Look at residual matrix (convert to matrix?) Check that entries match.

Week 5

Found bug. (In variances). Also note that diriclet variances should be between 0 and 1.

Check with $\gamma = 1$.

Check with λ

Vocab

- γ: step size
- ρ: phylogenetic distance scaling
- ω: correlation weighting.
- β estimated parameter values/ coefficients
- λ: diagonal scalar to subtract from Hessian.

Tuesday: Noticed code was getting ugly and hard to navigate. Changed return values to initialize easier by just initilizeing one list that has sub lists inside it.

Note: changed difference (plus convergence criteria) to be sum of square differences between beta between loop.

Wednesday: Trying on 10p data (by accident oops). Still takes a long time. Does not have large gaps. Omega approaches 1.

When running on the flavo data, it 'converges' in only 2 iterations. Also check out.

Reminder:

- $\omega = 1$ means all weight is compositional structure. If this is the case it doesnt really matter what ρ is

- $\omega = 0$ means all weight is phylogenetic structure.
- ρ close to 0 means:
- ρ large means:

Working on a function that outputs all diagnoistics plots. Done. Currently has phi, rho, omega, diffs. Need to add residuals information and beta information. how to visualize? Maybe a line plot with a line for each beta?

Trying 30p dataset which originally had decent values for omega and p. Quit in 2 iterations. Omega flipped.

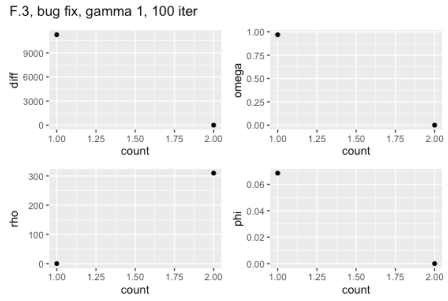

Figure 1:

Next step: step through function with 30p.

```
  Browse[2]> beta
 [1]  32.6820041 -35.2771786  -9.3769864   9.8059738  -6.3851386   1.6309199  -9.0146528
 [8]   5.1724875  -7.5425704   2.6913815  -9.4761818   4.5570750  -7.8544924   3.1674555

Browse[2]> as.vector(update)
 [1] -1.252986e-27 -1.252986e-27 -2.654175e-38 -2.654175e-38 -3.051694e-40 -3.051694e-40
 [7] -8.706320e-39 -8.706320e-39 -1.811488e-39 -1.811488e-39 -1.424226e-39 -1.424226e-39
[13] -7.777419e-40 -7.777419e-40 -4.486093e-40 -4.486093e-40 -3.262656e-40 -3.262656e-40
```

Seems like the 'update' is essentially zero.

Which is because the hessian is essentially zero.

```
  Browse[2]> as.vector(hess)
   [1]  -1.000000e-02  -9.860304e-46   1.007366e-76  -2.793269e-79  -2.314020e-64
   [6]  -2.492956e-64   3.834566e-57  -3.314361e-61  -2.560319e-62  -2.595960e-62
  [11]   1.079445e-60  -4.290110e-64  -3.059416e-64  -3.316491e-64   3.727629e-59
  [16]  -2.020623e-63  -6.455299e-63  -6.567340e-63  -1.639493e-49  -1.639493e-49
  [21]   1.620392e-65  -4.701085e-77   1.780107e-59  -2.344796e-60  -3.144451e-56
```

But estimating equations are also basically zero.

```
  Browse[2]> summary(as.vector(esteq))
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1.253e-29  0.000e+00  0.000e+00  0.000e+00  0.000e+00  1.253e-29
```

Wondering if this is a step thing? Try with gamma = .1 like before. Seems like a step thing. With gamma .1: Some WEIRD behavor with beta plot... Oh, this was some weird result of the beta plotting??? (Had X axis as descrete, but now it is not.)

Quick iteration seems to be a step problem. But betas are different For example on the dataset 30p, For gamma = 1 vs gamma = .1

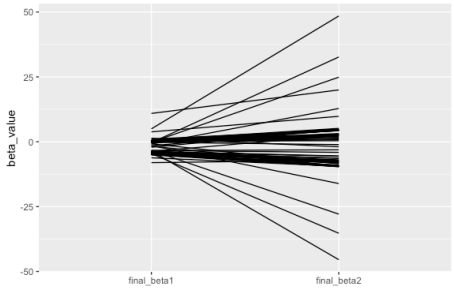

Figure 2:

.1 Independence design

Try setting R to be the identity matrix. Only need to change line in `calculate_equations` function.

Using values of gamma = .1, 100 iterations.

VERY DIFFERENT... Here final_beta2 is independence, final_beta1 is standard R.

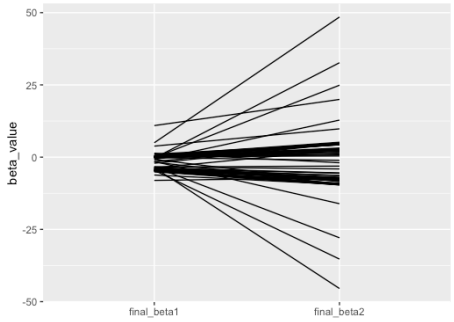

Figure 3:

Not sure what I do besides say: they are different!

Try to focus on one thing at once to be able to add git commits.

Shouldnt they be the same though? isnt idea that its just precision that changes? Am confused.

I think the important next step is to test independence design against pre-made functions? like geem. Can i actually use another package? Since family and link are weird. Look more into GEEM.

... Wait is my link wrong? Since isnt it logit? Or is it because I am linking alpha instead of mu?

Meeting

Bring up: What was the thing about the metabolomic data?

I explored model running performance after finding the bug. Found that the model "converges" after 2 iterations now, but if step size is reduced, does not. But gives different beta values.

Simulations

Start simulations

Idea that correlation matrix is redundant for dirichlet and we are making it not so

Week 6

Starting simulations!

Need: Example distance matrix. Either from zebrafish, or an example dataset in phyloseq? Using Global patterns, fecal samples. Randomly select (15?) taxa from more abundant taxa. Extract distance matrix.

Give some beta value. Should be for each taxa. The same? Different? Ask in email.

Split X as even 0s and 1s. Do i set an intercept?? No?

Set omega, set rho.

Calculate alpha as $\log \alpha = X\beta$.

Calculate R from α, ρ, D, ω.

Simulate $y \sim N(\alpha/\alpha_0, R)$ Which shares the same mean and covariance as the model. but Doesnt have the same variance? or does it?

Also R is covariance matrix not variance? So there is a σ^2 value?

Stop getting confused about variance/ covariance/ correlation.

Standardize. Ask? Calculate sum above zero etc, proportion... Report.

UMM I think the get_dirichlet_cor function is the CoVARIANCE. But for R i want the VARIANCE?? Nope it is ok. Acutally calculated the correlation and it is correct. The numerator looks funny.. Wait... actually there is missing a square root! Nope it is there in the code. Everything is fine.

$$\begin{aligned} Cor_D &= \frac{Cov}{\sqrt{V_i}\sqrt{V_j}} \\ &= -\frac{\alpha_i\alpha_j}{\alpha_0^2(\alpha_0+1)}\sqrt{\frac{\alpha_0^2(\alpha_0+1)}{\alpha_i(\alpha_0-\alpha_i)}}\sqrt{\frac{\alpha_0^2(\alpha_0+1)}{\alpha_j(\alpha_0-\alpha_j)}} \\ &= -\sqrt{\frac{\alpha_i\alpha_j}{(\alpha_0-\alpha_i)(\alpha_0-\alpha_j)}} \end{aligned}$$

NOTE! in rho omega fun, changed matrix to byrow = T. bug? test.

Hmm. Does this matter? it didnt change the beta values at all..

Final beta results

```
  [[50]]
 [1]   4.5424926  -2.3861832  -6.2525309   3.2004029   0.5698044  -3.6470294  -5.9934911
 [8]   3.1636061   5.8906071  -8.1679282  -6.1703515   4.7639107  -1.6003967   0.5118927
[15]  -6.1924151   4.6084286  -7.9448292 -10.0329420  -6.3725543   4.8713606   2.3284297
[22]  -3.8093811  -6.1630714   4.7997512   1.6008044  -3.3514829  -6.2586418   4.6867781
[29]  -2.6127452   0.7888159
```

Want

umm still getting infinities with this data..

Meeting

Run with only dirichlet instead (set omega = 1)

Make sure GEE runs with just dirichlet data. Try also with true correltion/covariance matrix V just to get right betas.

Seems like intercept term did not initialize well. Just have it be 0.

CHECK that matrix colsumns has correlctly byrow = T or byrow = F in it! make sure it is right.

Week 7

Trying to do better with version control. Did some reorganizing. See todo and commits

Want to work though to make sure indeces of everything are what I'm intending.

Do I make everything a dataframe with the intended index?

So do we allow an intercept even when simulated with no intecept? Yes?

Dirichlet simulation

Found one issue that might be why it wasnt working in the meeting. Alphas were being generated for each p not each n.

Comment out the wRrhores function in meeting.

Code in a way to only have dirichlet correlation.

Stepping through function: n = 100, p = 15, phi = 1, q = 2 (adds intercept), rho = 5 (doesnt matter since omega = 1 - all dirichlet)

Ok, ran through but still didnt work... Also the difference was (increasing???)

Now step through. Since we know the 'true' alpha values, we can compare.

Noticed update beta step still needed R inv. changed.

iter 0: all 30 beta values 0.

Seems like first step always has a better beta value. Do I have a wrong minus sign? ugh.

Ok, after "fixing" the X matrix:

try with no intercept: Betas only -2 If run for 20 iterations, get error in SVD error.... What is this caused by? why? Probably should add that diagonal? But since we have way more samples shouldnt be a problem right? Ok, upped n to 500 with no error.

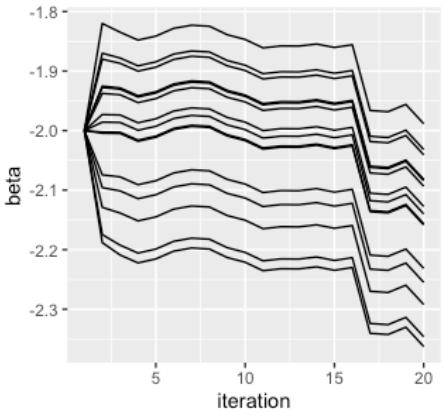

Figure 4:

OK, now try some betas as 2 and some as -1

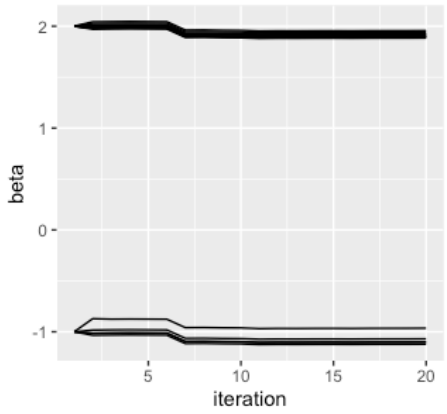

Figure 5:

OK

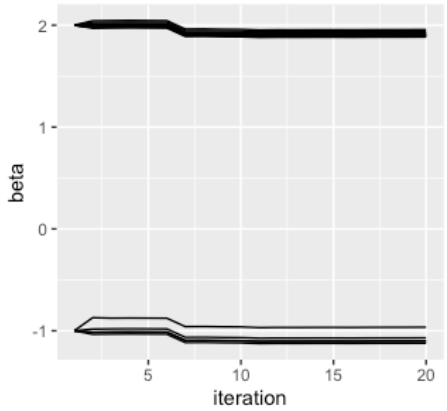

Figure 6:

WITH INTERCEPT

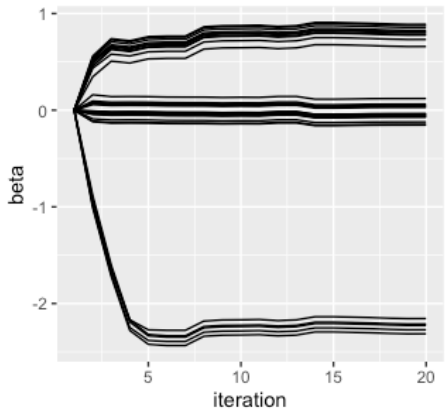

Figure 7:

Research Meeting

- Find out why didnt work this time
- Dirichlet matrix will be nonsingular so still include diagonal
- Rerun, and push to github changes... do the freakin versino control!
- Comment that variance structure is just dirichlet, but ok, since phylogenetic correlation is just correlation, not variance
- Check that this is correlation! right? not covariance. Probably correct since diagonal is 1.
- Phi shoudl be close to 1 since phi is 1 for dirichlet
- Literature review of hypothesis test

Ok, working for 0 and 3... (well if we have it start at 0 and 3)

Starts at 0 and 3:

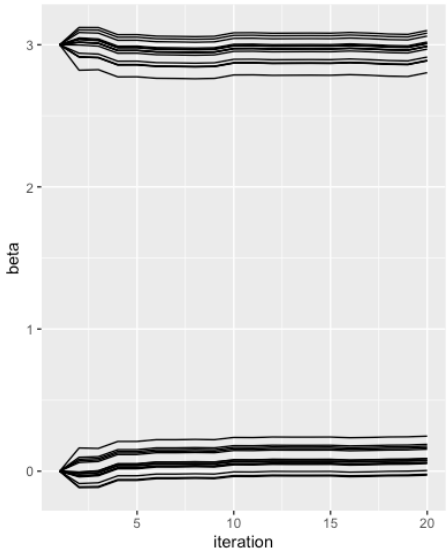

Figure 8:

Starts all at 0

Is this the number of iterations? Try up to 100 Ok, this is all 3, and 0 intercept... I dont even know...

Remove intercept? All 3 with no intercept, initialize at 0

Start with correct values, Stays at correct values

Wondering if actually wouldnt work for every otu to have a positive beta since we have negative correlation. add lambda to hessian:

TODO: Switch from dirichlet and see if it works normally!!!! Just like normal or poisson or something...

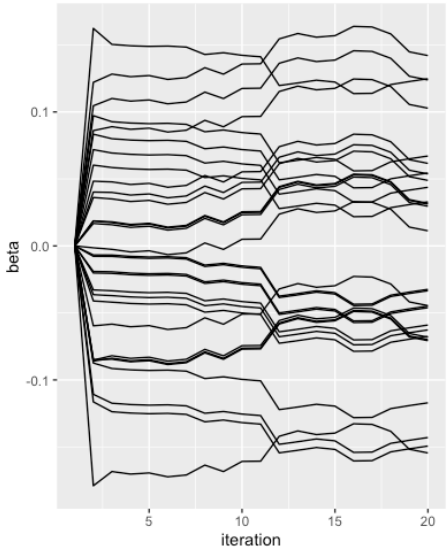

Figure 9:

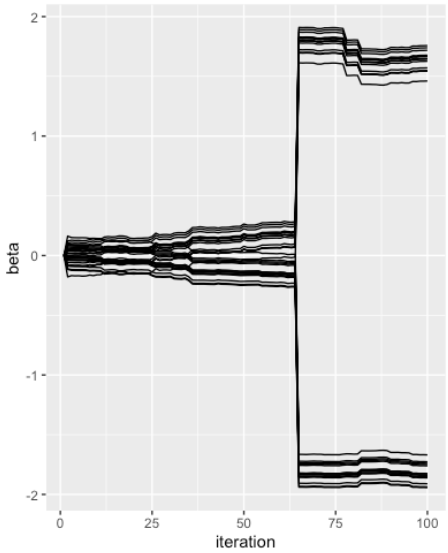

Figure 10:

Week 10

Todo:

Get a standard GEE working: Ex: normal distribution? poisson?

Meeting

- Identifiabilitly problem:
- A shift of betas doesnt affect mu....
- Try with continuous X

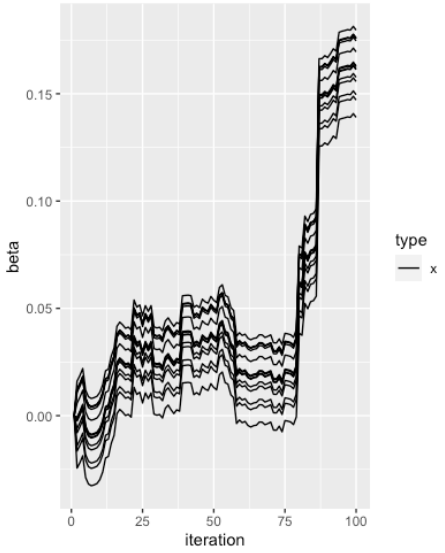

Figure 11:

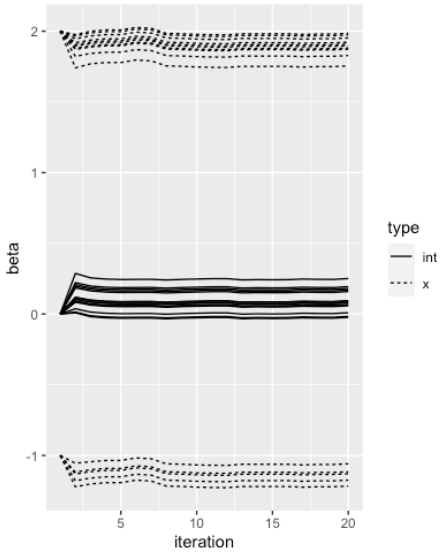

Figure 12:

- Maybe try $logit(\frac{\alpha_{ik}}{\alpha_{i0}})$
- Model the first k-1 then...
- Calculate partials, V, mean diff on true model, compare to estimator. print out those values.
- Suspician is that beta values can shift and wont change GEE values, will be close to 0.
- partials: $diag(mu_i) - mu_i mu_i^t$
- could also constrain beta, to be 0.
- G is derivative of quasiliklihood function. can have:

$$-q(\beta)+]lambda(\sum]beta_k)^2$$

- add derivative of this penalty to G.

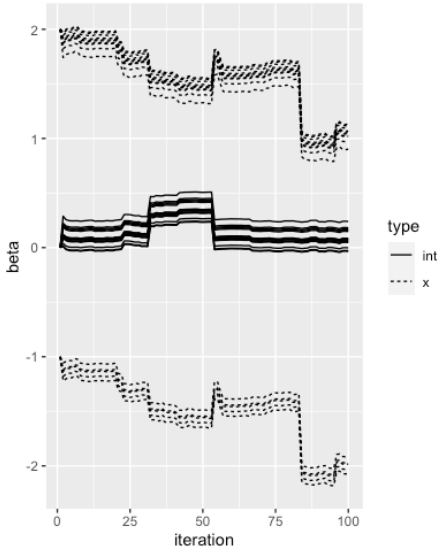

Figure 13:

- and add second derivative to hessian, fix.
- λ really big.
- quasi penalty: $\lambda\beta^t 11^t \beta$
- First dirivitive: $11^t\beta$
- second derivative: $2\lambda 11^t$
- Simulate with true sum is zero, and tru sum is not zero (betas)
- how does shift in parametrs change variance
-