Note from Alex: *Included is my data analysis draft, a figure, a table and some blue comments which may or may not be useful in discussion. I did the table two ways - the same information is contained in both versions (The* `mutlirow` *package is needed for the first one.) I'm happy to edit text, figures or tables if that's helpful.*

1 Data Analysis

1.1 Estimating microbial richness in Lake Champlain

To illustrate the performance of our methods on ecological data, we estimate strain-level microbial diversity in Lake Champlain, a large eutrophic lake in Canada. We analyze data from **?**, considering samples from the littoral zone in the summer season of the same year as replicates. This gives us 8 replicates from 2009, 6 replicates from 2010 and 6 replicates from 2011.

The richness estimates from each method are displayed in Figure 1 and Table 1. Method 0 produces the highest estimate for 2009 ($\widehat{C}_{[0]} = 73,404$) and 2010 ($\widehat{C}_{[0]} = 47{,}631$) and it was second highest to Method 2 in 2011 ($\widehat{C}_{[0]} = 57,686$ and $\widehat{C}_{[2]} = 118,547$). Method 1 consistently estimated C to be at or near the observed richness, $\widehat{C}_{[1]} = 584$ in 2009, 572 in 2010 and 718 in 2011. Methods 3 and 4, which both use goodness of fit, were between these two extremes. These results mirror what we observed in simulations, where Methods 0 and 2 generated the highest estimates and Method 1 was the lowest (see Figure **??**).

These estimates are up to 165 times the observed richness, suggesting that a majority of the species were unobserved in each sample. The relatively large estimates suggest that these data may be poorly fit by the negative binomial model.

Notes for discussion:

- In order to apply our methods to real data, some manual tuning of the optimization search grids (λ^{grid} and the search grid for C) was required.
- $\widetilde{\lambda}$ was similar for each method in 2010 and 2011 but larger in 2009, suggesting that $\widetilde{\lambda}$ may increase with r.
- The observed richness is an underestimate of the actual richness and the motivation for penalization is that the unpenalized MLE tends to overestimate the actual richness. Our data analysis shows that only the methods making use of a goodness of fit

metric (3 and 4) were consistently between Method 0 and the observed richness in all samples. We conclude that Methods 3 and 4 show the greatest promise on real data containing replicates.

Table 1: Diversity estimates from the Lake Champlain data analysis from 2009 ($r = 8$), 2010 ($r = 6$) and 2011 ($r = 6$) using our proposed methods.

Year	Method	$\widehat{C}$	$\widetilde{\lambda}$	$\widehat{\alpha}$	$\widehat{\delta}$
2009	[0] Unpenalized MLE	73,404	—	0.00088	0.00180
	[1] Minimum Variance	584	700	0.28863	0.00326
	[2] CV Likelihood	25,930	220	0.00516	0.00142
	[3] Goodness of fit	20,160	550	0.00323	0.00174
	[4] CV G.O.F.	36,893	85	0.00246	0.00251
2010	[0] Unpenalized MLE	47,631	—	0.00185	0.00253
	[1] Minimum Variance	572	500	0.62668	0.00724
	[2] CV Likelihood	24,799	0	0.00379	0.00270
	[3] Goodness of fit	13,156	225	0.00685	0.00257
	[4] CV G.O.F.	47,098	0	0.00208	0.00277
2011	[0] Unpenalized MLE	57,686	—	0.00161	0.00140
	[1] Minimum Variance	718	500	0.40112	0.00355
	[2] CV Likelihood	118,547	0	0.00122	0.00193
	[3] Goodness of fit	40,040	230	0.00231	0.00137
	[4] CV G.O.F.	36,395	5	0.00358	0.00178

Table 2: Diversity estimates from the Lake Champlain data analysis from 2009 ($r = 8$), 2010 ($r = 6$) and 2011 ($r = 6$) using our proposed methods.

	2009				2010				2011			
Method	$\widehat{C}$	$\widetilde{\lambda}$	$\widehat{\alpha}$	$\widehat{\delta}$	$\widehat{C}$	$\widetilde{\lambda}$	$\widehat{\alpha}$	$\widehat{\delta}$	$\widehat{C}$	$\widetilde{\lambda}$	$\widehat{\alpha}$	$\widehat{\delta}$
[0] Unpenalized MLE	73,404	—	0.00088	0.00180	47,631	—	0.00185	0.00253	57,686	—	0.00161	0.00140
[1] Minimum Variance	584	700	0.28863	0.00326	572	500	0.62668	0.00724	718	500	0.40112	0.00355
[2] CV Likelihood	25,930	220	0.00516	0.00142	24,799	0	0.00379	0.00270	118,547	0	0.00122	0.00193
[3] Goodness of fit	20,160	550	0.00323	0.00174	13,156	225	0.00685	0.00257	40,040	230	0.00231	0.00137
[4] CV G.O.F.	36,893	85	0.00246	0.00251	47,098	0	0.00208	0.00277	36,395	5	0.00358	0.00178

Figure 1: Estimates of strain-level microbial diversity in Lake Champlain in the summers of 2009 ($r = 8$), 2010 ($r = 6$) and 2011 ($r = 6$) using our proposed methods.