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 Tromas *and others* (2017), 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. These estimates are high as 160 times the observed richness, suggesting that the negative binomial model may be a poor fit to this data set. As a result, applying our methods to real data required some manual tuning of λ^{grid} and the search grid for C.

Despite the estimates being significantly more inflated than in our simulations, many of the trends observed in simulations hold true here. Method 0 produces the highest $\widehat{C}$ for 2009 and 2010 and it was second highest to Method 2 in 2011. Method 1 consistently estimated $\widehat{C}$ to be at or near the observed richness c. These are exactly the tendencies these methods showed when we generate simulations from a correctly specified model (see Figure **??**).

We model species richness because the observed richness is an underestimate of the truth. The original motivation for penalization is that the unpenalized MLE estimate can be much larger than the truth. In data analysis we don't know the truth, but we would expect an ideal method to be situated between these two extremes. Only the methods making use of a goodness of fit metric (3 and 4) are between Method 0 and the observed richness in all samples. As a result we conclude Methods 3 and 4 show the greatest promise on real data containing replicates.

Note from Alex: I've given two versions of the table, the first is much bigger and probably more suited to being a supplement. The second is more compact and could be crammed into a corner of the paper if needed. Happy to make revisions to tables and figures if that's helpful.

Table 1: Diversity estimates from the Lake Champlain data analysis from 2009 ($r = 8$), 2010 ($r = 6$) and 2011 ($r = 6$) from 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	593	700	0.27720	0.00342
	[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.	39,997	100	0.00578	0.00253
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	220	0.00685	0.00258
	[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

References

Tromas, Nicolas, Fortin, Nathalie, Bedrani, Larbi, Terrat, Yves, Cardoso, Pedro, Bird, David, Greer, Charles W and Shapiro, B Jesse. (2017). Characterising and predicting cyanobacterial blooms in an 8-year amplicon sequencing time course. *The ISME journal* **11**(8), 1746.

Table 2: Diversity estimates from the Lake Champlain data analysis from 2009 ($r = 8$), 2010 ($r = 6$) and 2011 ($r = 6$) from 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	593	700	0.27720	0.00342	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	220	0.00685	0.00258	40,040	230	0.00231	0.00137
[4] CV G.O.F.	39,997	100	0.00578	0.00253	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 ($r = 6$) based on our proposed methods.