

Tuning parameter selection for a penalized maximum likelihood  
estimator of species richness

Alexander Paynter

A thesis  
submitted in partial fulfillment of the  
requirements for the degree of

Master of Science

University of Washington

2019

Committee:

Amy Willis, Chair

James Hughes

Program Authorized to Offer Degree:  
Biostatistics

©Copyright 2019  
Alexander Paynter

University of Washington

**Abstract**

Tuning parameter selection for a penalized maximum likelihood estimator of species richness

Alexander Paynter

Chair of the Supervisory Committee:  
Amy Willis  
Department of Biostatistics

Our goal is estimating the true number of classes in a population. We focus on the scenario where multiple frequency count tables have been collected from the same population. In this setting we demonstrate the efficacy of a previously published penalized maximum likelihood method. Four novel methods to tune the requisite penalization parameter are proposed. The performance of all proposed tuning methods is compared in simulations.

# TABLE OF CONTENTS

	Page
Notation Summary . . . . .	1
Chapter 1: Introduction . . . . .	2
Chapter 2: Literature Review . . . . .	4
Chapter 3: Extension to Replicate Data . . . . .	8
Chapter 4: Efficacy of Penalization . . . . .	11
4.1 Simulation setup . . . . .	11
4.2 Simulation results . . . . .	13
Chapter 5: Methods for Tuning $\lambda$ . . . . .	16
5.0 Method 0: MLE without penalization . . . . .	16
5.1 Method 1: Minimum subset variance . . . . .	17
5.2 Method 2: Cross validated likelihood . . . . .	18
5.3 Method 3: Goodness of fit . . . . .	19
5.4 Method 4: Cross validated goodness of fit . . . . .	20
Chapter 6: Tuning Method Simulations . . . . .	21
6.1 Primary simulation . . . . .	21
6.1.1 Setup . . . . .	21
6.1.2 Results . . . . .	22
6.2 Secondary simulation . . . . .	25
6.2.1 Setup . . . . .	25
6.2.2 Results . . . . .	25
6.3 Simulation Conclusions . . . . .	27
Chapter 7: Discussion . . . . .	28
Chapter 8: Code References . . . . .	30

## **DEDICATION**

To my parents and grandparents  
for their steadfast support of my education.

## NOTATION SUMMARY

Symbol	Definition
$f_k$	The number of species observed $k$ times.
$X_i$	The number of times we observe an arbitrary species $i$ , $X_i \sim \text{Poisson}(\Lambda)$
$\Lambda$	Abundance parameter for an arbitrary species, $\Lambda \sim \text{Gamma}(\alpha, \delta)$
$\eta$	Abundance parameter shorthand, $\eta = (\alpha, \delta)$ .
$C$	$\sum_{k \geq 0} f_k$ , true or total species richness.
$c$	$\sum_{k \geq 1} f_k$ , observed species richness.
$p_\eta(x)$	Density for the gamma-Poisson distribution with parameter vector $\eta$ .
$k_{max}$	Largest $k$ such that $f_k > 0$ .
$f_{kj}$	The number of species observed $k$ times in replicate $j$ .
$X_{ij}$	The number of times we observe species $i$ in replicate $j$ .
$c_j$	$\sum_{k \geq 1} f_{kj}$ , observed richness in replicate $j$ .
$\{f_{kj}\}_{k \geq 1}$	Frequency count table for replicate $j$ .
$r$	The number of biological replicates in a sample.
$L(C, \eta)$	Likelihood (unpenalized)
$\ell(C, \eta)$	Log likelihood (unpenalized)
$\lambda$	Penalization constant (introduced in Wang and Lindsay (2005)).
$\lambda^{\text{grid}}$	The set of $\lambda$ values we iterate over in tuning proposals.
$J$	A subset of frequency count tables, $J \subseteq \{1, \dots, r\}$
$\mathcal{O}_\lambda(C, \eta; J)$	Objective function using $\lambda$ , replicate indices in $J$ .
$\widehat{C}_q$	The $C$ estimate from maximizing the objective function with $\lambda = q$ (Chapter 3).
$\widehat{C}_{[q]}$	The $C$ estimate obtained using method $q$ (Chapter 5).
$\widehat{\lambda}_{[q]}$	The $\lambda$ value selected by method $q$ .
$p$	The number of partitions.
$i$	$i \in \{1, \dots, C\}$ , the $i$ -th species
$j$	$j \in \{1, \dots, r\}$ , the $j$ -th replicate.
$k$	$k \in \{1, \dots, k_{max}\}$ , labels for the $f_k$ 's.
$l$	$l \in \{1, \dots, p\}$ , the $l$ -th partition.

## Chapter 1

### INTRODUCTION

In the *species problem* the goal is estimating the number of classes  $C$  which truly exist in a population. Typically exhaustive sampling is not possible, so our estimation is based on a sample of  $c < C$  classes. The problem is named for its origins in biological ecosystems, where  $C$  is *species richness*, the number of species. However, methods developed for the species problem can be readily applied to applications far removed from biology. For example, Efron and Thisted (1976) estimated the number of words  $C$  Shakespeare truly knew by considering the words he used in his works to be a sample of  $c$  classes. Fegatelli and Tardella (2018) estimated the number of cars  $C$  covered by an insurer based on accident data where  $c$  cars had at least one accident.

In ecology, species richness is worthy of estimation because it is one quantifiable aspect of ecosystem diversity. We focus our attention on microbial ecology where high species richness is associated with ecosystem stability, including in the human microbiome (see Sekirov et al. (2010)). Microbial data often contain many species observed only a few times (*rare species*) and some which are observed a large number of times (*abundant species*). In data with both rare and abundant species, richness estimation methods tend to have high variance.

Let  $f_k$  be the number of classes observed  $k$  times, called the *frequencies* or *frequency counts*. Then the set  $\{f_k\}_{k \geq 1}$  is a *frequency count table*, a common way to represent collected data. Using this notation, the species problem can then be rephrased as predicting  $f_0$  (always unobserved) because  $C = f_0 + f_1 + f_2 + \dots$  and  $c = C - f_0$ . In this paper we focus on the analysis of multiple frequency count tables. By demonstrating that the penalization proposal due to Wang and Lindsay (2005) is effective in analyzing data with replicates, we extend their results. Further, we show that replicate data is useful in tuning the required penalization

parameter.

This paper will be organized into the following chapters: We will review the literature most pertinent to our investigations in Chapter 2. In Chapter 3 we specify the extension to replicate data (more than one frequency count table). In Chapter 4 the utility of penalization in replicate analysis is established using simulations. In Chapter 5 we propose several novel methods for tuning the penalization parameter. In Chapter 6 we will evaluate our tuning proposals using simulations. Chapter 7 closes with an expanded discussion of our results and suggested directions for future work.

## Chapter 2

### LITERATURE REVIEW

Poisson mixture models were used in the earliest examinations of species abundance. In a Poisson model the number of times we observe an arbitrary species  $i$  is  $X_i \sim \text{Poisson}(\Lambda)$  where  $\Lambda$  is a random variable distributed according to some mixing distribution. The literature is filled with proposals for mixing distributions including the log normal in Bulmer (1974), inverse Gaussian in Ord and Whitmore (1986) and a nonparametric distribution composed of point masses proposed by Norris and Pollock (1998). In this paper we focus on the simple, easily interpretable gamma distribution, proposed to study species abundance by Fisher et al. (1943).

Let  $\Lambda \sim \text{Gamma}(\alpha, \delta)$  with distribution function  $f(\lambda) = \delta^\alpha \Gamma(\alpha)^{-1} \lambda^{\alpha-1} e^{-\delta\lambda}$ . We use  $\eta$  to reference the abundance parameters, so  $\eta = (\alpha, \delta)$  in the gamma-Poisson model. Greenwood and Yule (1920) showed that the distribution of  $X_i$  under a gamma-Poisson model can be written as a negative-binomial distribution,

$$p_\eta(x) = \frac{\Gamma(x + \alpha)}{\Gamma(\alpha)x!} \left( \frac{\delta}{1 + \delta} \right)^\alpha \left( \frac{1}{1 + \delta} \right)^x$$

which makes it an especially convenient model to work with.

There exist Bayesian (Efron and Tibshirani (1976)), objective Bayesian (Barger et al. (2010)) and other approaches to parameter estimation in addition to the maximum likelihood approach that Fisher used. However, we will focus on likelihood maximization in this paper. The most straightforward maximum likelihood approach is to simultaneously maximize  $C$  and  $\eta$  using the full likelihood, which is known as the *direct* approach. A complication in this model is that we have continuous parameters  $\eta$  as well as one discrete parameter  $C$ , so it is not immediately clear that derivative-based optimization methods are appropriate. This issue has been studied by Lindsay and Roeder (1987) and they provide a discrete analog to

the score function for this model and others. However, they point out that the asymptotics required use  $C \rightarrow \infty$ , which may not be appropriate for all problems.

There are two more major approaches to likelihood maximization, the *conditional* and *profile* approaches. In order to understand these we review the work of Sanathanan (1977), who showed we can decompose the likelihood for the gamma-Poisson model into two parts.

$$L(C, \eta) = L_b(C, \eta)L_c(\eta) \quad (2.1)$$

$$L_b(C, \eta) = \frac{C!}{(C-c)! c!} [1 - p_\eta(0)]^c [p_\eta(0)]^{C-c} \quad \text{Binomial likelihood.} \quad (2.2)$$

$$L_c(\eta) = \frac{c!}{\prod_{k \geq 1} f_k!} \prod_{k \geq 1} \left[ \frac{p_\eta(k)}{1 - p_\eta(0)} \right]^{f_k} \quad \text{Conditional likelihood.} \quad (2.3)$$

In this decomposition only  $L_b$  depends on  $C$ , while  $L_c$  is a function of  $\eta$  alone.  $L_b$  shows the probability of observing  $c$  species follows a binomial distribution, i.e.  $c \sim \text{Binomial}(C, 1 - p_\eta(0))$ . Letting  $k_{\max}$  be the largest  $k$  such that  $f_k > 0$ ,  $L_c$  shows that observing the frequency counts given by  $f_1, f_2, \dots, f_{k_{\max}}$  follows a multinomial distribution given  $c$ . The multinomial cell probabilities are proportional to  $p_\eta(k)$ .

Sanathanan (1977) uses this decomposition to propose the conditional approach to likelihood maximization, which is done in two steps (see Chao and Bunge (2002) for a detailed presentation). First we maximize  $L_c(\eta)$  to obtain a conditional abundance estimate  $\widehat{\eta}_c$ . Then  $L_b(C, \widehat{\eta}_c)$  is maximized over  $C$ . With  $\eta$  fixed there is a closed form expression for this second step,

$$\widehat{C}_c = \left\lfloor \frac{c}{1 - p_{\widehat{\eta}_c}(0)} \right\rfloor \quad (2.4)$$

where  $\lfloor a \rfloor$  is the largest integer less than or equal to  $a$ . In the *profile* likelihood approach the expression in Equation 2.4 is substituted into the full likelihood, which gives us a function of  $\eta$  alone (see Wang and Lindsay (2005)). The conditional and profile approaches have been shown to be asymptotically equivalent to the direct approach.

A perennial issue in the species problem is the instability of estimators, especially maximum likelihood estimators. The estimates often have unacceptably high variance and can

become dramatically inflated. Fisher noted that if there are many low abundance species then  $\alpha$  will be near 0. The existence of highly abundant species, as we see in microbial data, tends to cause  $\delta$  to be near 0 as well. The tendency of both abundance parameters to be near their boundary is one way to explain why this instability issue has plagued the species problem over a variety of approaches, and is especially prominent in microbial data.

One way of improving stability is to abandon the Poisson mixture paradigm entirely. Though we will focus on the gamma-Poisson model in this paper, we would be remiss if we did not briefly mention other popular alternatives. Coverage-based estimators, which use the proportion of species observed  $k$  times rather than the actual counts were originally proposed by Good (1953) and remain popular. Capture-recapture methods (see Chao (1987)) have a deep literature in animal abundance studies but are not as applicable in microbial data where the sampling destroys the organism. See Bunge and Fitzpatrick (1993) for a comprehensive review and Bunge et al. (2014) for an updated review of species richness methods which are outside the scope of this paper.

A proposal which retains the gamma-Poisson model and encourages stability is suggested by Wang and Lindsay (2005). Their proposal is to add a penalty term to the log likelihood, where the penalization is based on the probability of observing a species zero times (they give three penalty functions and we focus on only the first). If we let  $\ell(C, \eta)$  be the log likelihood, then they define a penalized log likelihood  $\ell_\lambda(C, \eta)$  as

$$\ell_\lambda(C, \eta) = \ell(C, \eta) - \lambda \log p_\eta(0) \tag{2.5}$$

where  $\lambda > 0$  is a penalization parameter. This encourages the maximum likelihood solution for  $\eta$  to fit so that  $p_\eta(0)$  is smaller. Therefore  $\alpha$  and  $\delta$  are further from the boundary, which alleviates some of the instability. The penalization term also causes a monotonic decrease in  $C$  estimates as greater values of  $\lambda$  are used. Consequently, if a large enough  $\lambda$  value is applied the  $C$  estimates will shrink to the observed richness  $c$ .

In their paper they show a trade-off exists: greater values of  $\lambda$  correspond to a more stable estimator, but at the potential cost of negative bias. This is a realization of the bias-variance

tradeoff. The choice of penalization parameter implies a preference for lower variance or lower bias, and therefore adds a degree of subjectivity to the estimation procedure (comparable to a Bayesian prior as discussed in Wang and Lindsay (2005)). We would also expect different data sets to require different  $\lambda$  values for optimal estimation, which makes direct application of this method to real data challenging. Our primary interest in studying the analysis of replicate frequency count tables is to leverage the replicates to tune the penalization parameter in Equation 2.5. The goal of this exploration is to obtain an estimator which is more stable than the unpenalized MLE but does not have the disadvantage of requiring a fixed, user-specified penalization parameter.

## Chapter 3

### EXTENSION TO REPLICATE DATA

Throughout we will focus on the gamma-Poisson model with penalization as proposed by Wang and Lindsay (2005). In this section we extend the penalized likelihood for a single frequency count table to the case where multiple frequency count tables have been collected.

Let  $r$  be the number of frequency count tables. We will assume that these  $r$  frequency count tables are biological replicates drawn independently from the same population, i.e. they have a common structure specified by the parameters  $C$  and  $\eta$ . A set of  $r$  frequency count tables from the same population will be called a *sample*, and a single frequency count table will be referred to as a *replicate*. Let the number of species observed  $k$  times in replicate  $j$  be  $f_{kj}$ ,  $j \in \{1, \dots, r\}$ . Let  $c_j$  be the observed richness in replicate  $j$ .

Microbial data is often available as labeled abundance information  $\{X_{ij}\}$ , the number of times species  $i$  was observed in replicate  $j$ . In this paper we study data represented in frequency counts  $\{f_{kj}\}$  and it is worth pointing out that we lose some information in this representation. For instance, in sample of frequency count tables it cannot be determined whether species  $i$  was observed in two different replicates. By using a mixed Poisson model we assume a common abundance structure between the species in order to estimate  $C$ . While there may be other methods which leverage the additional information in an abundance table representation, the frequency counts are a minimal sufficient statistic for mixed Poisson models.

In Chapter 5 we will require an objective function defined in terms of a subset of indices of the frequency count tables, so we define it that way now. Let  $J \subseteq \{1, \dots, r\}$ .  $J$  indicates the data being used in the evaluation of the objective function, with  $J = \{1, \dots, r\}$  meaning we use all replicates. Let  $\ell(C, \eta; \{f_{kj}\}_{k \geq 1})$  be the unpenalized log likelihood for replicate  $j$ .

Using the fact that the draws are independent, our objective function for a fixed  $\lambda \geq 0$  is the sum of the log likelihoods for each individual replicate. We will refer to this as the *penalized log likelihood* and it is given by

$$\mathcal{O}_\lambda(C, \eta; J) = \sum_{j \in J} \ell(C, \eta; \{f_{kj}\}_{k \geq 1}) - \lambda \log p_\eta(0) \quad (3.1)$$

$$= \sum_{j \in J} \left[ \log C! - \log(C - c)! + (C - c) \log p_\eta(0) - \sum_{k \geq 1} \log f_{kj}! + \sum_{k \geq 1} f_{kj} \log p_\eta(k) \right] - \lambda \log p_\eta(0) \quad (3.2)$$

By fixing  $\lambda = 0$  we obtain the *unpenalized log likelihood* for all replicates  $j \in J$ . We define  $\widehat{C}_\lambda$  to be the penalized maximum likelihood estimate using  $\lambda$ , that is

$$(\widehat{C}_\lambda, \widehat{\eta}_\lambda) = \arg \max_{C, \eta} \mathcal{O}_\lambda(C, \eta; \{1, \dots, r\}) \quad (3.3)$$

where  $\widehat{\eta}_\lambda$  is a nuisance parameter.  $\eta$  is critical to estimation but we are largely concerned with estimating  $C$  in the species problem.

We use root-mean-square error as the primary criterion for evaluating the performance of estimators in simulations. Let  $n_{\text{sim}}$  be the number of simulations completed for one choice of  $(C, \eta, r)$ , and  $\widehat{C}_{\lambda(s)}$  be the solution of Equation 3.3 obtained in simulation  $s$  with  $\lambda$ . Then we define

$$\text{RMSE}(\widehat{C}_\lambda) = \sqrt{\frac{1}{n_{\text{sim}}} \sum_{s=1}^{n_{\text{sim}}} (\widehat{C}_{\lambda(s)} - C)^2} \quad (3.4)$$

To maximize the likelihood we will use the direct approach, rather than the profile and conditional approaches reviewed in Section 2. While it may be faster and asymptotically equivalent to use the profile or conditional approach, our investigations revealed that the results are not always the same in finite samples. In this paper our end goal is to explore tuning, so the stability and accuracy of our underlying optimization algorithm have priority over speed. Crucially, will use exactly the same optimization for the likelihood throughout all simulations in this paper.

To implement a direct optimization approach we search over candidate  $C$  values in a grid, respecting the fact that  $C$  is a discrete parameter. A gradient-based search over  $\eta$  values is used at each  $C$  in the grid. The highest likelihood evaluation for any  $(C, \eta)$  is our maximum likelihood estimate. While this paper does not focus on optimization, our code is made available to the reader (see chapter 8 for links).

## Chapter 4

**EFFICACY OF PENALIZATION****4.1 Simulation setup**

Establishing that penalization can improve richness estimates is a crucial foundation before we attempt to evaluate tuning methods. Wang and Lindsay (2005) established that it can improve estimates for  $r = 1$  and in their particular application area. We need to show this for our models and  $r > 1$ . We propose to do so by simulating data under the gamma-Poisson model and maximizing the objective function with a set of fixed  $\lambda$  values. If there exists some  $\lambda > 0$  where  $\text{RMSE}(\widehat{C}_\lambda)$  over all simulations is less than  $\text{RMSE}(\widehat{C}_0)$  we conclude that penalization is effective under those simulation parameters.

We simulated using  $C = 1000$ , and  $r \in \{6, 10, 14\}$  to reflect the reality that few microbiome-focused labs are currently producing a large number of biological replicates. The abundance distribution of a population is characterized by  $\eta$ , so choosing appropriate simulation values of  $\eta$  is crucial in targeting our approach to microbial data. We used two microbial datasets as motivating examples. Walsh et al. (2014) provide a dataset from soil in a Swiss apple orchard and we found  $\eta = (\alpha, \delta) = (10^{-1}, 10^{-1})$  produced a frequency count table with a similar structure to their data. The second example from Tromas et al. (2017) is a sample of cyanobacteria from Lake Champlain, and we found  $\eta = (10^{-2}, 10^{-5})$  mimicked the structure of their data well. We note that is possible that the gamma-Poisson model is not the best fit for these examples, but our goal is only to choose  $\eta$  in an appropriate range. Two nearby  $\eta$  values are included,  $\eta = (10^{-1}, 10^{-3})$  and  $\eta = (10^{-1}, 10^{-5})$ . These add variety to our simulations in this section and context for our simulations in Section 6.2. Under each combination of  $\eta$  and  $r$  100 simulations were completed.

In Table 4.1 we provide a few summary measures for frequency count tables generated

**Table 4.1:** Summaries of the proportion of species which are unobserved, observed once (singletons), rarely observed and abundantly observed are calculated for each choice of  $\eta$ . We give the expected maximum abundance  $k_{\max}$ , the number of times the most abundant species was observed. In this table only  $\mathbb{E}(k_{\max})$  depends on  $C$ , and  $C = 1000$  is used here.

	$\eta =$	$(10^{-1}, 10^{-1})$	$(10^{-2}, 10^{-5})$	$(10^{-1}, 10^{-3})$	$(10^{-1}, 10^{-5})$
Proportion unobserved ( $p_{\eta}(0)$ )		0.787	0.891	0.501	0.316
Proportion singletons ( $p_{\eta}(1)$ )		0.072	0.009	0.050	0.032
Proportion rare ( $\sum_{k=1}^3 p_{\eta}(k)$ )		0.130	0.016	0.097	0.061
Proportion abundant ( $\sum_{k=10}^{\infty} p_{\eta}(k)$ )		0.028	0.083	0.340	0.583
Expected max abundance ( $\mathbb{E}[k_{\max}]$ )		$4.1 \times 10^1$	$2.0 \times 10^5$	$4.0 \times 10^3$	$3.9 \times 10^5$

under each choice of  $\eta$ . This characterizes the  $\eta$  values by the proportion of unobserved, rare or abundant species we would expect to see. We make this information available here for the interested reader but delay our discussion of the differences until Chapter 6 where they strongly influence our results.

One final setting is necessary for these simulations: a set of  $\lambda$  values to test over. Our primary concern is to ensure the set covers the optimal value, i.e.  $\lambda$  which produces the lowest  $\text{RMSE}(\widehat{C}_{\lambda})$ . While we do not have a theoretical way to determine the ideal  $\lambda$  set, we can use the known behavior of the penalty function to confirm that the set was sufficiently large after the simulation is complete. As  $\lambda$  increases,  $\widehat{C}_{\lambda}$  decreases. If  $\widehat{C}_{\lambda_0}$  is below the true  $C$ , then for  $\lambda > \lambda_0$   $\widehat{C}_{\lambda}$  will be further from  $C$ , and the RMSE will increase. Then we can be assured that the set is large enough if most or all simulations had a  $\widehat{C}_{\lambda}$  below the truth for the largest  $\lambda$  tested. This is clearly verifiable in the cases presented in Figures 4.1 and 4.2. Simulating using  $\lambda \in \{0, 5, \dots, 120\}$  was sufficient for all simulations in this section.

**Table 4.2:** Simulation results for various abundance ( $\eta$ ) and replicate ( $r$ ) choices. For each row in the table 100 simulations were run, e.g. for the first row we generated a group of 6 frequency count tables under  $C = 1000$ ,  $\eta = (10^{-1}, 10^{-1})$  one hundred times.  $\lambda_{opt}$  is the optimal value of  $\lambda$ , i.e.  $\lambda$  which produced the lowest RMSE over all simulations.  $\widehat{C}_0$  is the  $C$  estimate with  $\lambda = 0$ , and  $\widehat{C}_{\lambda_{opt}}$  is the  $C$  estimate using  $\lambda = \lambda_{opt}$ .

$\eta = (\alpha, \delta)$	$r$	RMSE ( $\widehat{C}_0$ )	$\lambda_{opt}$	RMSE ( $\widehat{C}_{\lambda_{opt}}$ )
$(10^{-1}, 10^{-1})$	6	796.90	10	326.50
$(10^{-1}, 10^{-1})$	10	527.53	15	286.27
$(10^{-1}, 10^{-1})$	14	337.31	10	235.51
$(10^{-2}, 10^{-5})$	6	735.95	15	516.37
$(10^{-2}, 10^{-5})$	10	700.21	20	456.04
$(10^{-2}, 10^{-5})$	14	666.55	20	470.35
$(10^{-1}, 10^{-3})$	6	200.09	20	156.29
$(10^{-1}, 10^{-3})$	10	213.22	55	142.64
$(10^{-1}, 10^{-3})$	14	243.42	55	148.26
$(10^{-1}, 10^{-5})$	6	401.17	55	147.13
$(10^{-1}, 10^{-5})$	10	283.91	25	137.04
$(10^{-1}, 10^{-5})$	14	415.19	70	126.72

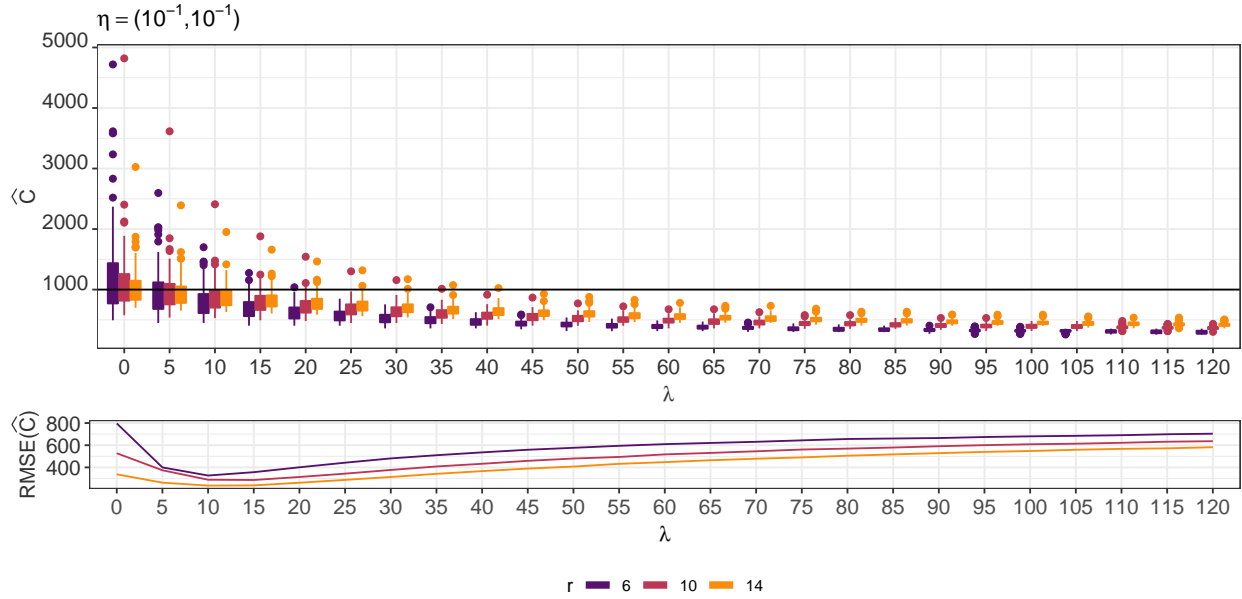
## 4.2 Simulation results

The simulation parameters and RMSE results are shown in Table 4.2. For every parameter choice a reduction in  $\text{RMSE}(\widehat{C}_\lambda)$  is attained for some  $\lambda > 0$ . Depending on simulation parameters the reduction varied from 22% to 70% compared to  $\text{RMSE}(\widehat{C}_0)$  for the best  $\lambda$ . These results show that for a variety of  $(\eta, r)$  choices penalization improves species richness estimates. In other words, if we knew the optimal  $\lambda$  value we could improve our richness estimates using penalization with that  $\lambda$  applied. This extends the results of Wang and Lindsay (2005) to the analysis of replicate frequency count tables and our particular application. We also note that depending on the simulation settings the optimal  $\lambda$  value varies considerably,

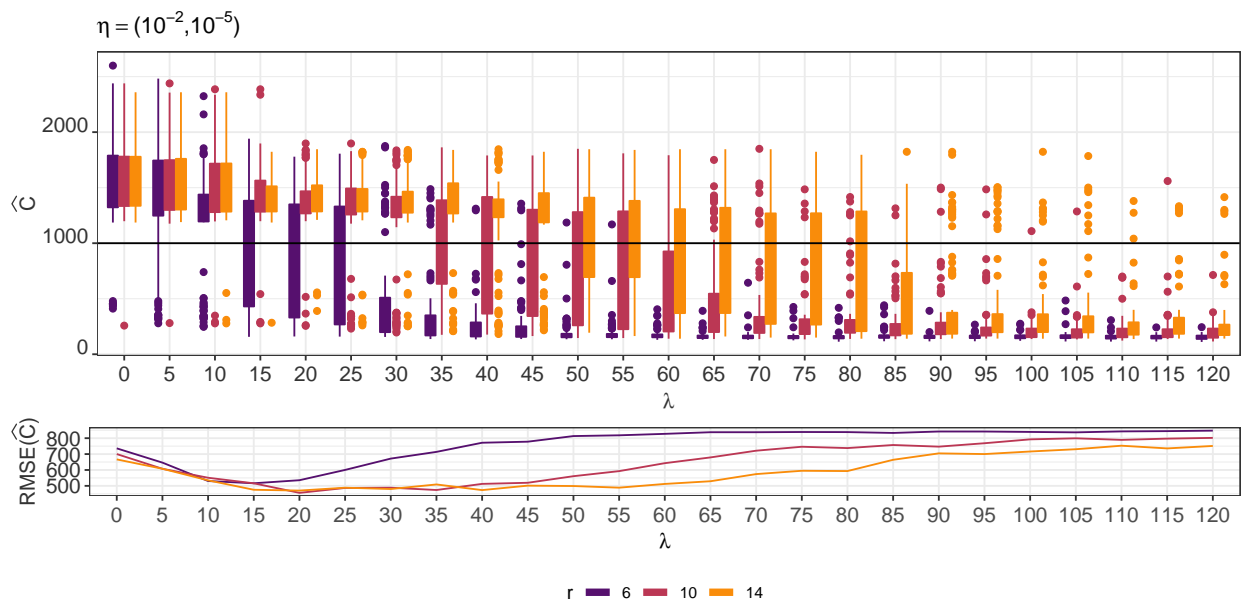
ranging from 10 to 70. This simulation suggests that there is not a universally appropriate  $\lambda$  choice. Therefore, we have evidence that tuning the  $\lambda$  value to the sample is highly desirable.

In order to demonstrate the effect on  $\widehat{C}_\lambda$  as we increase  $\lambda$ , we plot the results for two  $\eta$  values in Figures 4.1 and 4.2. For small values of  $\lambda$  we acquire lower variance in the estimates while incurring some negative bias, and the RMSE improves as a result. Eventually the bias difference dominates and the RMSE increases as  $\lambda \rightarrow \infty$ .

**Figure 4.1:** Simulation results for estimation of  $C$  under a range of fixed  $\lambda$  constants,  $\eta = (10^{-1}, 10^{-1})$ , over 100 simulations. The black line indicates the true value  $C = 1000$ .



**Figure 4.2:** Simulation results for estimation of  $C$  under a range of fixed  $\lambda$  constants,  $\eta = (10^{-2}, 10^{-5})$ , over 100 simulations. The black line indicates the true value  $C = 1000$ .



## Chapter 5

METHODS FOR TUNING  $\lambda$ 

We have established that penalization can improve richness estimates, but different abundance structures ( $\eta$ ) require different penalization parameters ( $\lambda$ ) to minimize RMSE. Therefore a method which can tune the parameter based on a sample is warranted. We propose several novel and intuitive methods in this chapter. Each is evaluated in simulations in Chapter 6.

**5.0 Method 0: MLE without penalization**

We begin by proposing a reasonable baseline to compare our tuning methods with. The easiest extension to the replicate case is to maximize the unpenalized likelihood using  $J = \{1, \dots, r\}$ . Recall that fixing  $\lambda = 0$  gives us the unpenalized likelihood in Equation 3.2, so if we denote the  $C$  estimate obtained using Method 0 as  $\widehat{C}_{[0]}$ , then

$$\left(\widehat{C}_{[0]}, \widehat{\eta}_{[0]}\right) = \arg \max_{C, \eta} \mathcal{O}_0(C, \eta; \{1, \dots, r\}) \quad (5.1)$$

This method is faster and makes fewer assumptions than any other method, so it serves as the standard to beat.

For all of the remaining methods (1-4) we generate estimates  $\widehat{C}_\lambda$  over  $\lambda \in \lambda^{\text{grid}}$ , where  $\lambda^{\text{grid}}$  is a user specified constant required to run each method (e.g.  $\lambda^{\text{grid}} = \{0, 5, 10, \dots, 60\}$  in Section 6.1). The methods differ from one another in how they subset the data to create these estimates, and in the steps they take to arrive at a final estimate  $\widehat{C}_{[\text{method}]}$ .

### 5.1 Method 1: Minimum subset variance

Method 1 is motivated by the belief that if we resample from the same population, an ideal  $C$  estimator should have low variance. Exploiting the fact that we have replicate data, the idea is to repeatedly partition the replicates into two subsets and come up with two estimates. We select the  $\lambda$  which yields the lowest between-subset variance. This partitioning is repeated  $p$  times to average out the arbitrary choice of subsets.

Let  $T_1(l)$  be the first subset of the  $l^{\text{th}}$  partition and  $T_2(l)$  be the second subset of the  $l^{\text{th}}$  partition, for  $l \in \{1, \dots, p\}$  ( $T$  is for training). We select the subsets randomly and we are using the word partition in the mathematical sense. So  $T_1(l), T_2(l) \subseteq \{1, \dots, r\}$ ,  $T_1(l) \cap T_2(l) = \emptyset$  and  $T_1(l) \cup T_2(l) = \{1, \dots, r\}$ . For each  $\lambda \in \lambda^{\text{grid}}$ ,  $l \in \{1, \dots, p\}$  let

$$\left(\widehat{C}_\lambda^{T_1(l)}, \widehat{\eta}_\lambda^{T_1(l)}\right) = \arg \max_{C, \eta} \mathcal{O}_\lambda(C, \eta; T_1(l)) \quad (5.2)$$

$$\left(\widehat{C}_\lambda^{T_2(l)}, \widehat{\eta}_\lambda^{T_2(l)}\right) = \arg \max_{C, \eta} \mathcal{O}_\lambda(C, \eta; T_2(l)) \quad (5.3)$$

We now have a  $\widehat{C}$  corresponding to each  $\lambda \in \lambda^{\text{grid}}$  in each subset of each partition. Our goal in Method 1 is to use these estimates to select the  $\lambda$  value which gave us the lowest average variance over all partitions, denoted by  $\widetilde{\lambda}_{[1]}$ . The overall estimate for Method 1,  $\widehat{C}_{[1]}$ , is a simple average of the estimates produced under  $\widetilde{\lambda}_{[1]}$ .

$$\widetilde{\lambda}_{[1]} = \arg \min_{\lambda} \frac{1}{p} \sum_{l=1}^p \text{Var} \left[ \widehat{C}_\lambda^{T_1(l)}, \widehat{C}_\lambda^{T_2(l)} \right] \quad (5.4)$$

$$\widehat{C}_{[1]} = \frac{1}{p} \sum_{l=1}^p \left[ \frac{\widehat{C}_{\widetilde{\lambda}_{[1]}}^{T_1(l)} + \widehat{C}_{\widetilde{\lambda}_{[1]}}^{T_2(l)}}{2} \right] \quad (5.5)$$

Practically speaking, in our simulations we use a 50/50 split of the indices for each partition. The subsets of each partition are selected at random, that is we could select the same subset split more than once. We partition a total 10 times ( $p = 10$ ). For example if  $r = 4$  then  $T_1(1) = \{1, 4\}$  and  $T_2(1) = \{2, 3\}$  would be a valid subset selection for partition 1. This holds in methods 2 and 4 as well, which also use partitioning.

## 5.2 Method 2: Cross validated likelihood

A well known method which takes advantage of replicate data is cross validation. Cross validation is most commonly applied to evaluate the strength of a model in a dataset containing both features and outcomes. The model is trained on one subset of the data and then evaluated using another. However, in the species problem we never observe true outcomes, i.e.  $C$  is always estimated using the frequency counts. Methods 2 and 4 are motivated by the desire to do something similar to cross validation, but we have to find a function which takes the place of mean prediction error in the absence of any true  $C$  observations.

In Method 2 we use one subset of the data to generate estimates. Then we evaluate those estimates using the likelihood with the data from the other subset. As before, we partition the data into two subsets  $p$  times, calling them  $T(l)$  for the training subset of the  $l$ -th partition, and  $E(l)$  for the evaluation subset of the  $l$ -th partition.  $T(l), E(l)$  are created exactly as in Method 1, we only relabel them to indicate the change in purpose. Then for each  $\lambda \in \lambda^{\text{grid}}$ ,  $l \in \{1, \dots, p\}$  let

$$\left(\widehat{C}_\lambda^{T(l)}, \widehat{\eta}_\lambda^{T(l)}\right) = \arg \max_{C, \eta} \mathcal{O}_\lambda(C, \eta; T(l)) \quad (5.6)$$

so that we have an estimate using only the training subset for each  $\lambda$  value and each partition. A  $\lambda$  value is selected by plugging those estimates into the unpenalized likelihood with  $J = E(l)$ .

$$\widetilde{\lambda}_{[2]} = \arg \max_\lambda \sum_{l=1}^p \mathcal{O}_0\left(\widehat{C}_\lambda^{T(l)}, \widehat{\eta}_\lambda^{T(l)}; E(l)\right) \quad (5.7)$$

$$\widehat{C}_{[2]} = \frac{1}{p} \sum_{l=1}^p \widehat{C}_{\widetilde{\lambda}_{[2]}}^{T(l)} \quad (5.8)$$

In the evaluation step given in Equation 5.7 we must use the unpenalized likelihood, as opposed to the penalized version. The penalty term is in fact always positive, and has the potential to dominate the likelihood if a very large  $\lambda$  exists in  $\lambda^{\text{grid}}$ . For example, suppose we used the penalized likelihood and  $10^7 \in \lambda^{\text{grid}}$ . Then no matter what data was passed in, Method 2 would select  $\widetilde{\lambda} = 10^7$ , and  $\widehat{C}_{[2]}$  would shrink to the observed richness as a consequence.

### 5.3 Method 3: Goodness of fit

We depart from our partitioning scheme to explore another desirable property of an estimator. An ideal estimator would produce parameter estimates  $\widehat{C}$ ,  $\widehat{\eta}$  which fit well with the frequency counts in our sample. Consider a single frequency count table. Using Equation 2.1 we can find the expected frequency counts  $f_k$  given  $C$  and  $\eta$ ,

$$\mathbb{E}[f_k] = \mathbb{E}[\mathbb{E}[f_k|c]] = \mathbb{E}\left[c \frac{p_\eta(k)}{1 - p_\eta(0)}\right] = C(1 - p_\eta(0)) \frac{p_\eta(k)}{1 - p_\eta(0)} = Cp_\eta(k) \quad (5.9)$$

Then given estimates  $\widehat{C}$  and  $\widehat{\eta}$  we use a plug-in estimate for the expected counts:  $\widehat{f}_k = \widehat{C}p_{\widehat{\eta}}(k)$ . If we consider each  $f_k$  a category, the usual  $\chi^2$  goodness of fit statistic is

$$\sum_{k=1}^{\infty} \frac{(f_k - \widehat{f}_k)^2}{\widehat{f}_k} \quad (5.10)$$

$$= \sum_{k=1}^{\infty} \frac{(f_k - \widehat{C}p_{\widehat{\eta}}(k))^2}{\widehat{C}p_{\widehat{\eta}}(k)} \quad (5.11)$$

$$= \sum_{k=1}^{k_{max}} \frac{(f_k - \widehat{C}p_{\widehat{\eta}}(k))^2}{\widehat{C}p_{\widehat{\eta}}(k)} + \widehat{C} \sum_{k=k_{max}+1}^{\infty} p_{\widehat{\eta}}(k) \quad (5.12)$$

Equation 5.12 uses the fact that we have some  $k_{max}$  in every frequency count table such that  $f_k = 0$  for all  $k > k_{max}$ . Therefore the statistic is a convergent sum, because the infinite sum is the tail of a distribution multiplied by a finite number  $\widehat{C}$ . Equation 5.12 is useful in software implementation as we can make use of built-in CDF functions for the tail. Nonetheless we use Equation 5.11 below in extending to replicates for more concise notation.

In method 3 we make use of this goodness of fit metric by first generating estimates using all replicates. For each  $\lambda \in \lambda^{\text{grid}}$  let

$$(\widehat{C}_\lambda, \widehat{\eta}_\lambda) = \arg \max_{C, \eta} \mathcal{O}_\lambda(C, \eta; \{1, \dots, r\}) \quad (5.13)$$

$\widehat{C}_{[3]}$  is determined by evaluating the goodness of fit metric shown above for each choice of  $\lambda$ .

$\tilde{\lambda}_{[3]}$  is the  $\lambda$  value which provides  $\widehat{C}_\lambda$  estimates which fit the data best.

$$\tilde{\lambda}_{[3]} = \arg \min_{\lambda} \sum_{j=1}^r \sum_{k=1}^{\infty} \frac{(f_{kj} - \widehat{C}_\lambda p_{\widehat{\eta}_\lambda}(k))^2}{\widehat{C}_\lambda p_{\widehat{\eta}_\lambda}(k)} \quad (5.14)$$

$$\widehat{C}_{[3]} = \widehat{C}_{\tilde{\lambda}_{[3]}} \quad (5.15)$$

In Equation 5.14 note the estimated counts  $(\widehat{C}_\lambda p_{\widehat{\eta}_\lambda}(k))$  are the same over all replicates ( $j$ ). This follows because we do not have independent  $C, \eta$  estimates from each replicate, but a shared estimate for the sample.

#### 5.4 Method 4: Cross validated goodness of fit

In Method 4 we return to the partitioning scheme of Method 2. Rather than using the likelihood in the evaluation step, we hypothesize that the goodness of fit metric may be a better choice. Partition the data  $p$  times into  $T(l)$ , the training set in partition  $l$ , and  $E(l)$ , the evaluation set in partition  $l$ . For all  $\lambda \in \lambda^{\text{grid}}$  and  $l \in \{1, \dots, p\}$ , we generate estimates exactly as in Method 2,

$$(\widehat{C}_\lambda^{T(l)}, \widehat{\eta}_\lambda^{T(l)}) = \arg \max_{C, \eta} \mathcal{O}_\lambda(C, \eta; T(l)) \quad (5.16)$$

To select  $\lambda$  we evaluate the goodness of fit metric using only the evaluation subset data,  $j \in E(l)$ . The  $\lambda$  which produces the best fitting estimates is  $\tilde{\lambda}_{[4]}$ ,

$$\tilde{\lambda}_{[4]} = \arg \min_{\lambda} \sum_{l=1}^p \sum_{j \in E(l)} \sum_{k=1}^{\infty} \frac{(f_{kj} - \widehat{C}_\lambda^{T(l)} p_{\widehat{\eta}_\lambda^{T(l)}}(k))^2}{\widehat{C}_\lambda^{T(l)} p_{\widehat{\eta}_\lambda^{T(l)}}(k)} \quad (5.17)$$

$$\widehat{C}_{[4]} = \frac{1}{p} \sum_{l=1}^p \widehat{C}_{\tilde{\lambda}_{[4]}}^{T(l)} \quad (5.18)$$

Compared to Method 2, this method generates estimates  $\widehat{C}_\lambda^{T(l)}$  in the same fashion but evaluates in a completely different way. Compared with Method 3, this method uses each replicate to either generate an estimate or evaluate the fit, while in Method 3 all the replicates are used in both steps.

## Chapter 6

### TUNING METHOD SIMULATIONS

We have proposed four tuning methods motivated by properties which would be desirable in an estimator of  $C$ . In this chapter we compare the RMSE of estimates from each method in simulations.

To evaluate which method has the best performance we will apply them to gamma-Poisson simulated data with the same parameters used in Chapter 4. Based on the results in Chapter 4 we know that estimation can be improved on this set of parameters through the use of a fixed penalization parameter. The purpose of this chapter is to determine if any method can reliably select a  $\lambda$  that improves the RMSE.

We present two simulation below, and in each it is important to select an appropriate  $\lambda^{\text{grid}}$ . In particular our grid needs to be large enough to reveal the negative aspects of methods which over-penalize. For instance, if  $\lambda^{\text{grid}} = \{0, \dots, 50\}$  and the optimal value is  $\lambda = 45$ , a method which deterministically selects the maximum value in  $\lambda^{\text{grid}}$  (an extreme over-penalizer) would perform well. This would lead us to spuriously conclude this hypothetical over-penalizer is effective. One of the important foundations laid in Section 4 is that we now know approximately what the optimal values are for each simulation setting (Table 4.2). We select a large enough grid by extending to approximately double the optimal value in both simulations presented.

#### 6.1 Primary simulation

##### 6.1.1 Setup

In this simulation we let  $C = 1000$ , and simulate 100 times over all combinations of  $\eta \in \{(10^{-1}, 10^{-1}), (10^{-2}, 10^{-5})\}$ ,  $r \in \{6, 10, 14\}$ . Recall these  $\eta$  are the values chosen based on

our motivating examples. We know that the optimal  $\lambda$  values are between 10 and 20 under these  $\eta$  parameters for  $r$  frequency count tables, so we chose  $\lambda^{\text{grid}} = \{0, 5, 10, \dots, 60\}$  for this simulation. Methods 0-4 are all evaluated over the same random draws for each parameter choice.

### 6.1.2 Results

In Table 6.1 we show the RMSE over all simulations for each method and each  $\eta, r$  combination. Over all parameter choices Method 3 performs at least as well as Method 0, with a RMSE which was 0-23% lower depending on  $\eta$  and  $r$ . Method 3 is the only method which is at least as good as Method 0 for all parameter choices.

Universally superior RMSE is the most desirable property for a method, but we can characterize the performance of the methods based on the simulation parameters as well. For example,  $\eta = (10^{-2}, 10^{-5})$  has more abundant species and larger expected maximum abundance  $\mathbb{E}(k_{\max})$ , but relatively few rare species (see Table 4.1). The RMSE for Method 2 under this  $\eta$  is 14-24% smaller than Method 0, and for Method 4 the RMSE is slightly better for 2 of the 3  $r$  values. Conversely,  $\eta = (10^{-1}, 10^{-1})$  has more rare species and fewer abundant species and it is only Method 3 which has consistently improved performance.

Method 1 has better performance only for  $r = 6$  on both  $\eta$  choices. In general, we would expect that any truly effective method would improve relative to Method 0 as  $r$  increases, as this means more data was collected. The fact that Method 1 loses its edge over Method 0 as we collect more data is a discouraging finding. Method 3 shows the opposite trend, which is especially clear in Figure 6.1 on  $\eta = (10^{-2}, 10^{-5})$ . As we increase  $r$  the simulations appear closer to the truth in Method 3, and Method 3 has the best RMSE for  $r = 14$  under both  $\eta$  values.

In Figure 6.1 we display the  $C$  estimates for each method. Recall that as we increase  $\lambda$ ,  $\widehat{C}_\lambda$  monotonically decreases. Therefore we can discuss whether a method is over-penalizing (selecting  $\widetilde{\lambda}$  which is larger than optimal) or under-penalizing (selecting  $\widetilde{\lambda}$  which is smaller than optimal) without additional information. If  $\widehat{C}$  is too high then the method has under-

**Table 6.1:** RMSE( $\widehat{C}_{[\text{method}]}$ ) over all simulations. Simulation parameters are grouped by  $\eta$  and  $r$ , with  $C = 1000$  fixed throughout. Under each  $\eta, r$  combination 100 simulations were run. Methods with RMSE better than Method 0 have a blue background, and the best method is **bolded**.

	$\eta = (10^{-1}, 10^{-1})$			$\eta = (10^{-2}, 10^{-5})$		
	$r = 6$	$r = 10$	$r = 14$	$r = 6$	$r = 10$	$r = 14$
Method 0: MLE (no penalization)	709	<b>326</b>	<b>339</b>	787	775	716
Method 1: Minimum subset variance	<b>689</b>	630	578	<b>763</b>	821	796
Method 2: Cross validated likelihood	797	521	492	<b>602</b>	<b>658</b>	617
Method 3: Goodness of fit	707	<b>326</b>	<b>339</b>	781	663	<b>554</b>
Method 4: Cross validated g.o.f.	812	571	533	<b>738</b>	787	679

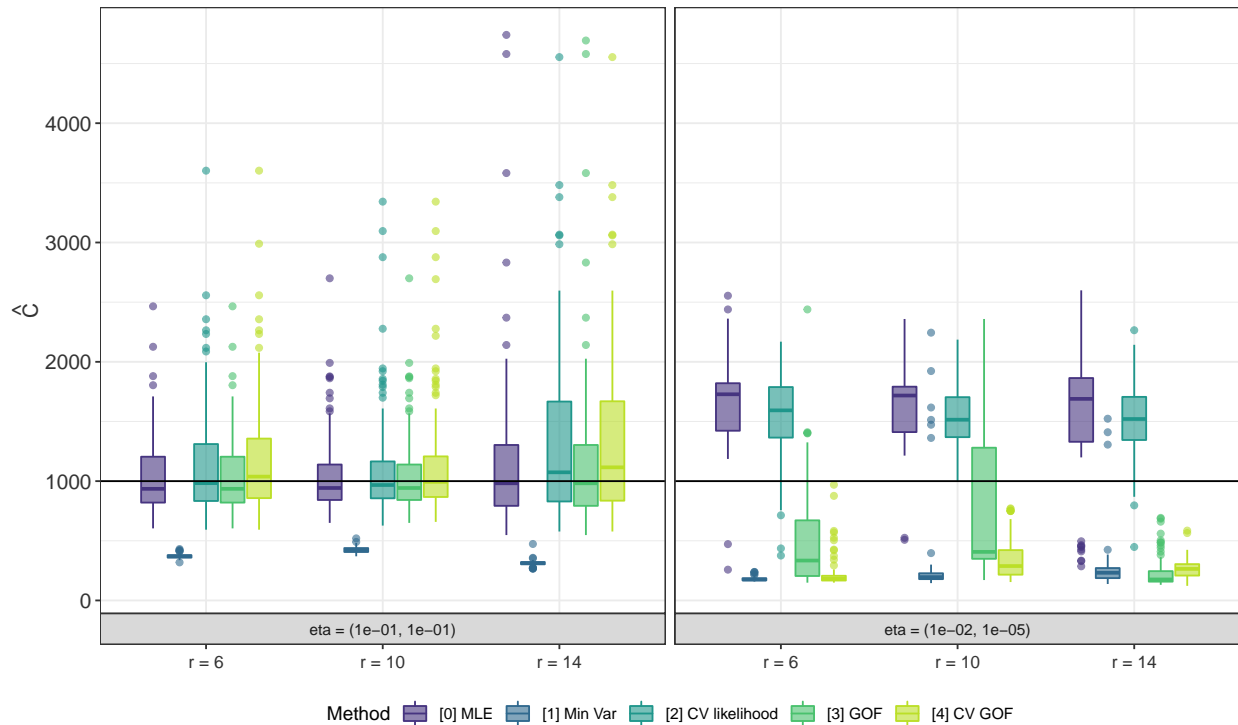
penalized and if  $\widehat{C}$  is too low then the method has over-penalized.

It is clear in Figure 6.1 that Method 1 over-penalizes for all parameter choices, as all of the estimates are far below the truth. We can make sense of this observation based on what we observed in Figure 4.1: for very high values of  $\lambda$  the estimates have low variance because they are driven down to the observed richness  $\max_j c_j$ . However the observed richness has a large negative bias. Method 1 inherits this large bias as it minimizes variance, so RMSE is not effectively minimized.

Method 2 has the opposite behavior: the estimates tend to be too high, so Method 2 underpenalizes. This is especially noticeable in  $\eta = (10^{-2}, 10^{-5})$ , where the Method 0 bias is substantial. Under this  $\eta$  value Method 2 does better than Method 0, but the estimates are almost all too high.

Even for  $\eta = (10^{-1}, 10^{-1})$ , where the Method 0 bias is smaller, we see that Method 2 has poor performance (Table 6.1). Recall that in each partition of Method 2 we use only half the data to generate the estimates  $\widehat{C}_\lambda^{T(l)}$ . As a consequence Method 2 has slightly higher variance when compared with Method 0, as evidenced by the interquartile range in Figure 6.1 for  $\eta = (10^{-1}, 10^{-1})$ . In conclusion, Method 2 does not seem to be effectively tuning  $\lambda$ , making

**Figure 6.1:** Simulation results for all proposed tuning methods over  $\eta = (10^{-1}, 10^{-1})$  and  $\eta = (10^{-2}, 10^{-5})$  are shown, with a black line indicating the true value,  $C = 1000$ .



the cost of using only half the data to generate estimates especially problematic.

Methods 3 and 4 have different penalization tendencies depending on the  $\eta$  and  $r$  values of the simulation. This is a desirable property because we know the optimal  $\lambda$  value depends on the structure of the data, and we want a method which can adapt. In Methods 3 and 4 we generate estimates using the penalized likelihood and evaluate the fit with the goodness of fit metric. Incorporating two distinct functions may explain why these methods have more adaptable behavior, as opposed to Method 2 which uses likelihood exclusively.

While we find the indications that Method 4 is adaptable to the dataset encouraging, under  $\eta = (10^{-1}, 10^{-1})$  the RMSE for Method 4 is 15-75% greater than for Method 0, while Method 3 is at least as good as Method 0. Method 3 is adaptive while being simpler, faster and having better performance than Method 4. Given deficits noted for Method 1 and 2,

**Table 6.2:** RMSE in  $\widehat{C}_{[\text{method}]}$ . Simulation parameters are grouped by  $\eta$  and  $r$ , with  $C = 1000$  fixed throughout. Under each  $\eta, r$  combination 100 simulations were run. If Method 3 has better performance than Method 0 is has a **blue** background and it is **bolded**. This is redundant but consistent with Table 6.1.

	$\eta = (0.1, 0.001)$			$\eta = (0.1, 0.00001)$		
	$r = 6$	$r = 10$	$r = 14$	$r = 6$	$r = 10$	$r = 14$
Method 0: MLE (no penalization)	<b>228</b>	<b>236</b>	<b>188</b>	<b>163</b>	426	681
Method 3: Goodness of fit	260	274	256	183	<b>175</b>	<b>191</b>

Method 3 is by far the most promising method based on this simulation. We propose to conduct a secondary simulation to further challenge Method 3, testing whether it remains superior to Method 0 over a wider range of  $\eta$  values.

## 6.2 Secondary simulation

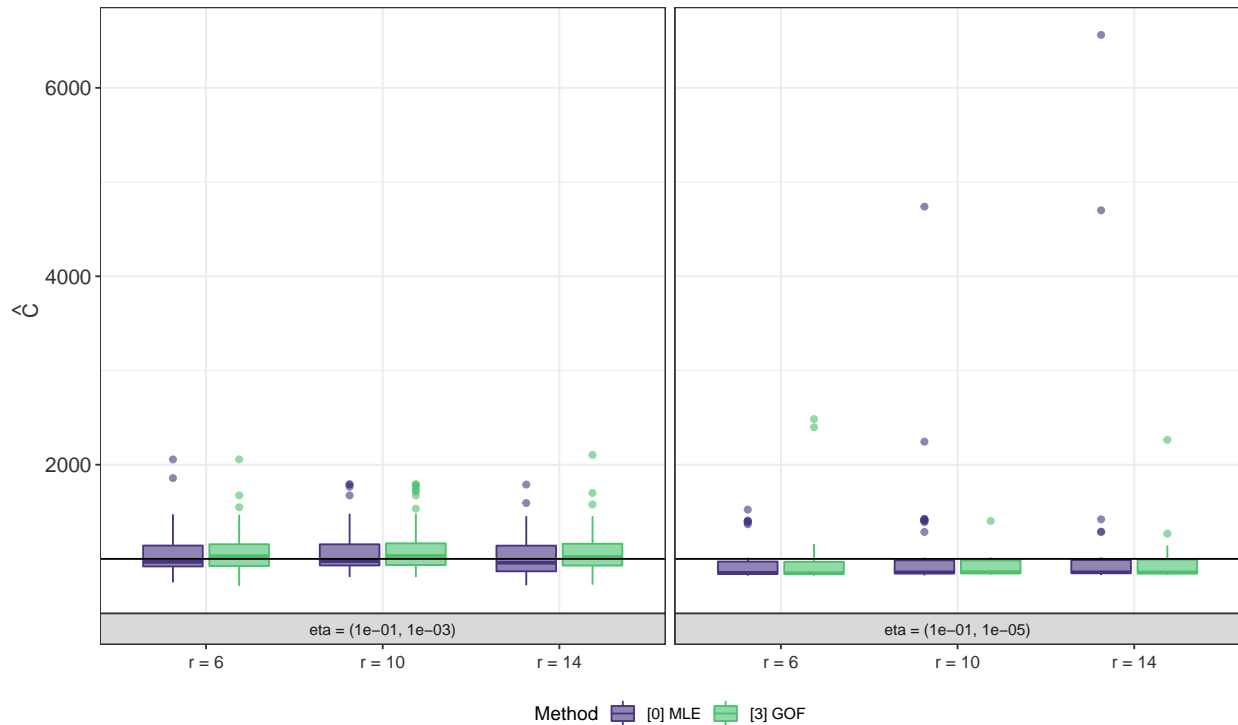
### 6.2.1 Setup

In this simulation we will fix  $C = 1000$ ,  $r \in \{6, 10, 14\}$  and do 100 simulations for each combination of  $\eta, r$  as before. We let  $\eta \in \{(10^{-1}, 10^{-3}), (10^{-1}, 10^{-5})\}$  and only consider methods 0 and 3. To account for the fact that a larger  $\lambda$  was optimal for these  $\eta$  values (see Table 4.2) we use  $\lambda^{\text{grid}} = \{0, 10, 20, \dots, 140\}$ .

### 6.2.2 Results

In Table 6.2 we display the RMSE for both methods and each  $\eta, r$  combination. Method 3 has slightly worse performance on all  $r$  values for  $\eta = (10^{-1}, 10^{-3})$ , with a RMSE which was 14-36% greater than Method 0. This  $\eta$  generates more rare species so the relatively poor performance of Method 3 is consistent with the results our previous simulation. In Section 6.1, Method 3 had greatly improved estimation in data with many abundant species, but the improvements over Method 0 were weaker in datasets that had many rare species.

**Figure 6.2:** Simulation results for methods 0 and 3 over  $\eta = (10^{-1}, 10^{-3})$  and  $\eta = (10^{-1}, 10^{-5})$  are shown, with a black line indicating the true value,  $C = 1000$ .



We also note that Method 3 has 12% higher RMSE on  $r = 6, \eta = (10^{-1}, 10^{-5})$ , but as  $r$  increases Method 3 improves relative to Method 0. This finding is also consistent with our previous simulation where Method 3 tends to perform better as more data is introduced.

In Figure 6.2 we display boxplots of the  $C$  estimates over all simulations. For  $\eta = (10^{-2}, 10^{-5})$  we note that  $\widehat{C}_{[0]}$  is noticeably inflated on a few simulations. Referencing Table 4.1 we can recall this  $\eta$  choice has the highest proportion of abundant species (while still having some singletons and unobserved species). Therefore this is an excellent example of a dataset which will cause the instability problem in maximum likelihood estimation we reviewed in Chapter 2. Method 3 succeeds on this  $\eta$  primarily by selecting lower estimates on these problematic simulations.

### **6.3 Simulation Conclusions**

As a result of these simulations, we conclude that none of the methods (including method 0) are best for all simulation settings. Method 1 resulted in a strong negative bias no matter which data are being analyzed. Method 3 had better performance than Method 0 when there are highly abundant species present in the data, which makes Method 0 highly unstable. When there are many rare species Method 3 is better than methods 1, 2 and 4 but not necessarily Method 0. Method 3 also improved relative to Method 0 with greater  $r$  and appears to better adapt to the data. Methods 2 and 4 had lower RMSE than Method 0 when highly abundant species exist, but we cannot recommend them over Method 3 due to their inferior performance on data dominated by rare species.

## Chapter 7

### DISCUSSION

In this paper we have proposed an extension of the penalization in Wang and Lindsay (2005) to data with biological replicates and proposed several methods for tuning the penalization parameter. This extends the existing body of literature on species richness by demonstrating that penalization is useful in analyzing replicate data. Our goodness of fit tuning method (Method 3) has better performance than the unpenalized MLE in 8 of the 12 parameter combinations we analyzed. In this section we note promising directions for future work.

Our results from Chapter 3 hint at a positive correlation between  $r$  and the optimal choice of  $\lambda$ . We hypothesized that this may be the case and considered other ways of extending the objective function given in Equation 3.2. For instance, using  $-h(J)\lambda \log p_\eta(0)$  where  $h(J)$  is  $|J|$  or  $\sqrt{|J|}$  seemed like plausible extensions, whereas the definition we give uses  $h(J) = 1$ . Our investigations revealed that the trend is unfortunately not this simple as  $|J|$  becomes large. A parsimonious extension which accounts for this observed trend would be a valuable modification to our work.

In Wang and Lindsay (2005) the gamma-Poisson model is explicitly considered but their results apply to many other Poisson mixture models. Our most promising method, Method 3, is also amenable to other modeling choices provided that we have some way to calculate the fitted counts  $\hat{f}_k$ . It would be interesting to explore whether our results hold for a more flexible mixed Poisson distribution, such as the proposal by Norris and Pollock (1998). Flexible models often have better performance under model misspecification, and we simulated under correctly specified models. Therefore, this variation on Method 3 would be especially interesting to study under model misspecification.

We intentionally chose a likelihood optimization algorithm which was stable and exhaustive but admittedly slow. Confirming that our results hold when a faster approach is used would be another valuable extension to our work. We did not derive theoretical standard errors in this paper, and the long computation times precluded the consideration of a bootstrapping approach. However, if our results hold when a faster likelihood maximization algorithm is used then a standard error procedure could become feasible.

We have been careful to try to create methods which only depend on  $\lambda^{\text{grid}}$  being large enough to contain the optimal value, with no change in results if it is too large. This makes  $\lambda^{\text{grid}}$  a more forgiving parameter than fixed  $\lambda$  values, but we still need to make a choice to analyze data. Faster computation times would aid in this problem because we could set a default grid which is generously large, thus alleviating concerns that our results are dependent on our choice. Additionally, it may be possible to select  $\lambda^{\text{grid}}$  for the methods adaptively. For instance, we could expand the grid until the fixed estimates (e.g.  $\widehat{C}_\lambda$  in Equation 5.13) have been suppressed to the observed richness. Choosing  $\lambda^{\text{grid}}$  remains a consideration when applying our methods.

Method 3 was our top performer and it is the only tuning method we proposed which can be run without replicates ( $r = 1$ ). The ever-decreasing cost of high-throughput sequencing suggests biological replicates might be more available in the future. However, there are currently more projects with a single frequency count table available. Exploring the use of Method 3 for single frequency count tables is a worthwhile extension of our work for this reason.

## Chapter 8

**CODE REFERENCES**

Our method implementations including our core objective optimization algorithm can be found at [github.com/statdivlab/rre](https://github.com/statdivlab/rre). The code to reproduce our figures and simulations can be found at [github.com/statdivlab/rre\\_sims](https://github.com/statdivlab/rre_sims). All code is written in R (R Core Team (2018)).

## BIBLIOGRAPHY

- BARGER, KATHRYN, BUNGE, JOHN ET AL. (2010). Objective bayesian estimation for the number of species. *Bayesian Analysis* **5**(4), 765–785.
- BULMER, MG. (1974). On fitting the poisson lognormal distribution to species-abundance data. *Biometrics*, 101–110.
- BUNGE, JOHN AND FITZPATRICK, MICHAEL. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association* **88**(421), 364–373.
- BUNGE, JOHN, WILLIS, AMY AND WALSH, FIONA. (2014). Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application* **1**, 427–445.
- CHAO, ANNE. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 783–791.
- CHAO, ANNE AND BUNGE, JOHN. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics* **58**(3), 531–539.
- EFRON, BRADLEY AND THISTED, RONALD. (1976). Estimating the number of unseen species: How many words did shakespeare know? *Biometrika* **63**(3), 435–447.
- FEGATELLI, DANILO ALUNNI AND TARDELLA, LUCA. (2018). Moment-based bayesian poisson mixtures for inferring unobserved units. *arXiv preprint arXiv:1806.06489*.
- FISHER, RA, CORBETT, AS AND WILLIAMS, CB. (1943). The relationship between the number of species and the number of individuals in a random sample of an animal population: *Journal animal ecology*, v. 12.

- GOOD, IRVING J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40**(3-4), 237–264.
- GREENWOOD, MAJOR AND YULE, G UDNY. (1920). An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal statistical society* **83**(2), 255–279.
- LINDSAY, BRUCE G AND ROEDER, KATHRYN. (1987). A unified treatment of integer parameter models. *Journal of the American Statistical Association* **82**(399), 758–764.
- NORRIS, JAMES L AND POLLOCK, KENNETH H. (1998). Non-parametric mle for poisson species abundance models allowing for heterogeneity between species. *Environmental and Ecological Statistics* **5**(4), 391–402.
- ORD, J KEITH AND WHITMORE, G ALEX. (1986). The poisson-inverse gaussian distribution as a model for species abundance. *Communications in Statistics-theory and Methods* **15**(3), 853–871.
- R CORE TEAM. (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- SANATHANAN, LALITHA. (1977). Estimating the size of a truncated sample. *Journal of the American Statistical Association* **72**(359), 669–672.
- SEKIROV, INNA, RUSSELL, SHANNON L, ANTUNES, L CAETANO M AND FINLAY, B BRETT. (2010). Gut microbiota in health and disease. *Physiological reviews* **90**(3), 859–904.
- 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.

WALSH, FIONA, SMITH, DANIEL P, OWENS, SARAH M, DUFFY, BRION AND FREY, JÜRIG E. (2014). Restricted streptomycin use in apple orchards did not adversely alter the soil bacteria communities. *Frontiers in microbiology* **4**, 383.

WANG, JI-PING Z AND LINDSAY, BRUCE G. (2005). A penalized nonparametric maximum likelihood approach to species richness estimation. *Journal of the American Statistical Association* **100**(471), 942–959.