Bayesian hierarchical modeling of size spectra

Jeff S. Wesner, Justin P.F. Pomeranz, James R. Junker, Vosjava Gjoni, Yuhlong Lio

University of South Dakota, Department of Biology, Vermillion, SD 5706

Colorado Mesa University

Louisiana University Marine Consortium

Michigan State University

University of South Dakota, Department of Mathematics, Vermillion, SD 57069

Jeff.Wesner@usd.edu

Abstract

A fundamental pattern in ecology is that smaller organisms are more abundant than larger organisms. At the individual level, this pattern is known as the individual size distribution (ISD) or community size spectrum, which describes the frequency distribution of all individuals in an ecosystem, regardless of taxon. This distribution arises from a power law distribution, such as the Pareto distribution, and a major goal of size spectra analyses is to estimate the ISD power law exponent. However, while numerous methods have been developed, they have focused almost exclusively on estimating the exponent from single samples. Here, we develop an extension of the truncated Pareto distribution within the probabilistic modeling language Stan. We use it to estimate multiple ISD exponents simultaneously with a hierarchical modeling approach. The most important result is the ability to examine hypotheses related to size spectra within a single generalized (non)-linear mixed model.

Keywords: *Bayesian, body size spectra, hierarchical, Pareto, power law, Stan*

Introduction

In any ecosystem, large individuals are typically more rare than small individuals. This fundamental feature of ecosystems generates a remarkably common pattern in which relative abundance declines with individual body size, generating the individual size distribution ISD (or community size spectrum) (Sprules and Barth 1983; White et al. 2008). More formally, the ISD is a frequency distribution that can be approximated by a bounded power law with a single free parameter λ, corresponding to the following probability density function (Edwards et al. 2020):

$$f(x) = Cx^{\lambda}, x_{min} \leq x \geq x_{max}$$

where x is the body size (e.g., mass or volume) of an individual regardless of taxon, x_{min} is the smallest individual attainable and x_{max} is the largest possible individual (White et al. 2008). C is a constant equal to:

$$C = \begin{cases} \frac{\lambda+1}{x_{max}^{\lambda+1} - x_{min}^{\lambda+1}}, \lambda \neq -1 \\ \frac{1}{log x_{max}^{\lambda+1} - log x_{min}^{\lambda+1}}, \lambda = -1 \end{cases}$$

This model is also known as the bounded power law or truncated Pareto distribution. The terms "bounded" or "truncated" refer to the limits of x_{min} and x_{max}. Without those limits, the function is a simple power law. Each term in the equations above comes directly from the data except for the exponent λ, which must be estimated from a statistical model.

The exponent λ is often described by its "steepness", with more negative values (i.e., "steeper") indicating higher abundance of large relative to small individuals, and vice versa. These patterns of size frequency are an emergent property of demographic processes (e.g., age-dependent mortality), ecological interactions (e.g., size-structured predation, trophic transfer efficiency), and physiological constraints (e.g., size-dependent metabolic rates) (Andersen and Beyer 2006, White et al. 2008). As a result, variation in λ across ecosystems or across time can indicate fundamental shifts in com-

munity structure or ecosystem functioning. For example, overfishing in marine communities has been detected using size spectra in which λ was steeper then expected, indicating fewer large fish than expected (Blanchard et al. 200x). Shifts in λ have also been used to document responses to acid mine drainage in streams (Pomeranz et al. 2019; Pomeranz et al. 2020), and changes in response to land use (Martínez et al. 2016), resource subsidies (Perkins et al. 2018), and temperature (Pomeranz et al. 2022).

Given the ecological information it conveys, the data required to estimate size spectra are deceptively simple; only a single column of data are needed, in which each data point is a single measure of the body size of an individual. As long as the body sizes are collected systematically and without bias towards certain taxa or phenotypes, there is no need to know any more ecological information about the data points (e.g., taxon, trophic position, age, abundance). However, despite the simple data requirement, the statistical models used to estimate λ are diverse. Edwards et al (2017) documented 8 different analytical methods. Six of these involved binning, in which the body sizes are grouped into size bins (e.g., 2-49 mg, 50-150mg, etc.) and then counted, generating values for abundance within each size bin. This allows λ to be estimated using simple linear regression. Unfortunately, the binning process also removes most of the variation in the data, collapsing information about 1000's of individuals into just 6 or so bins. Doing so can lead to the wrong values of λ, sometimes drastically so (White et al. 2008, Edwards et al. 2017/2020).

An improved alternative to binning and linear regression is to fit the body size data to a power law probability distribution (White et al. 2008, Edwards et al. 2017/2020). This method uses all of the data without binning and directly estimates λ using maximum likelihood (Edwards et al. 2017). However, at present it is only possible to fit the models to single data set using maximum likelihood. To our knowledge there is no current method to fit ISD models to multiple groups of ISD's, such as data collected from multiple sites or multiple years. Instead, hypothesis testing with individual size distributions is typically done in two steps (e.g., Pomeranz et al. 2022). First, λ estimates are obtained individually from each collection (e.g., each site or year, etc.). Then, these estimates are used as response variables in a linear model to examine how they relate to predictor variables

4

70 (Edwards et al. 2020). A downside to this approach is that it treats λ values as independent samples,

71 even if they come from the same sites or times. It also removes information on sample size (number

72 of individuals used to derive λ. As a result, the approach not only separates the data generation

73 model from the predictor variables, but it is also unable to take advantage of partial pooling during

74 model fitting.

75 Here, we develop a Bayesian model that uses the truncated Pareto distribution to estimate λ in

76 response to both fixed and random predictor variables. The model extends the maximum likeli-

77 hood approach developed by Edwards et al. (2020) and allows for a flexible hierarchical structure,

78 including partial pooling, within the modeling language Stan (Stan Development Team 2022).

79 **Methods**

80 *Translating to Stan*

81 We first translated the probability density function described by Edwards et al. (2020) into Stan by

82 converting it to a log probability density function (lpdf). Stan is a probabilistic modeling language

83 that is capable of fitting complex models, including those with custom lpdf's. The resulting lpdf is

84 below:

$$
C = \begin{cases} \log \frac{\lambda+1}{x_{max}^{\lambda+1} - x_{min}^{\lambda+1}} + \lambda \log x, \lambda \neq -1 \\ \log(\log x_{min} - \log x_{max}) + \lambda \log x, \lambda = -1 \end{cases}
$$

85 which is coded in Stan as follows:

```
functions{

real paretocustom_lpdf(real x, real b, real xmin, real xmax){

if(b != -1)
```

```
return(log((b+1) / ( xmax\^(b+1) - xmin\^(b+1))) + b\*log(x));



else



return(log(log(xmin) - log(xmax)) + b\*log(x));

    }



}
```

86   We call this the $paretocustom$ distribution, which we can now use to estimate λ of given a data

87   set. For example, an intercept-only model would look like this:

$$x_i \sim paretocustom(\lambda, x_{min}, x_{max})$$

88

$$\lambda = \alpha$$

89

$$\alpha \sim Normal(\mu, \sigma)$$

90   where x_i is the ith individual body size, λ is the size spectrum exponent, x_{min} and x_{max} are as

91   defined above, and α is the intercept with a prior probability distribution. In this case, we specified

92   a Normal prior since λ is continuous and can be positive or negative, but this can be changed as

93   needed.

94   The simple model above can be expanded to a generalized linear mixed model by including fixed

95   predictors ($\beta\mathbf{X}$) and/or varying intercepts ($\alpha_{[x]}$):

$$x_{ij} \sim paretocustom(\lambda_j, x_{min,j}, x_{max,j})$$

96

$$\lambda = \alpha + \beta\mathbf{X} + \alpha_{[j]} + \alpha_{[x]}$$

6

$$\alpha \sim Normal(\mu_\alpha, \sigma_\alpha)$$

$$\beta \sim Normal(\mu_\beta, \sigma_\beta)$$

$$\alpha_{[j]} \sim Normal(0, \sigma_{[j]})$$

$$\sigma_{[j]} \sim Exponential(\phi)$$

$$\alpha_{[x]} \sim Normal(0, \sigma_{[x]})$$

$$\sigma_{[x]} \sim Exponential(\phi)$$

with one or more β regression parameters β for one or more fixed predictors **X**, and one or more varying intercepts α_x. We specify α_j separately because it is needed to account for the non-independence of body sizes. In other words, each body size x_i is clustered within each site and so they are not independent and identically distributed. The addition of a varying intercept for each sample accounts for this non-independence. We demonstrate how excluding the sample-specific varying intercept can lead to overdispersion below. Prior distributions are given as $Normal$ for the parameters and varying intercept and $Exponential$ for $\sigma[x]$, but these can also be changed as needed.

The model above assumes that each body size x represents a single individual such that the data set might have many repeats for individuals of the same size (e.g., $x = \{0.2, 0.2, 0.2, 0.4, 0.4, 0.5, 9.8\}$). However, when individaul body sizes are repeated in a data set, they are often accompanied by a count or density, such that the data set above might instead consist of two columns with $x = \{0.2, 0.4, 0.5, 9.8\}$ and $counts = \{3, 2, 1, 1\}$. To analyze this more compact data set, Edwards et al. (2020) developed a modification of the log probability density function to include $counts$:

$$C = \begin{cases} counts(\log\frac{\lambda+1}{x_{max}^{\lambda+1}-x_{min}^{\lambda+1}} + \lambda\log x), & \lambda \neq -1 \\ counts(\log(\log x_{min} - \log x_{max}) + \lambda\log x), & \lambda = -1 \end{cases}$$

which is coded in Stan as follows:

```stan
functions{
  real paretocounts_lpdf(real x, real lambda, real xmin, real xmax, real counts){
    if(lambda != -1)
    return(counts*(log((lambda+1) / ( xmax^(lambda+1) - xmin^(lambda+1))) + lambda*log(x
    else
    return(counts*(log(log(xmin) - log(xmax)) + lambda*log(x)));
  }
}
```

We refer to this as $paretocounts$, such that the model can be fit using:

$$x_i \sim paretocounts(\lambda, x_{min}, x_{max}, counts)$$

$$\lambda = [\text{linear or non-linear model}]$$

$$[\text{priors}]$$

Aside from adding $counts$, the model is the same as presented above. These models ($paretocustom$ and $paretocounts$) allow us to test how the size distribution exponent, λ, varies in response to continuous or categorical predictors and to include hierarchical structure as needed.

Testing the models

The $paretocustom$ and $paretocounts$ lpdf's give the same results, differing only in how the data are aggregated. For simplicity, we demonstrate model performance here for the $paretocounts$ distribution, since the empirical data we used (see *Case Study* below) contains counts of individual body sizes. First, we tested for parameter recovery using data simulated from a bounded power law

129 with known values of λ. Second, we fit the model to fisheries trawl data presented in Edwards et

130 al. (2020) to estimate the hypothesis that λ declines over time.

131 *Parameter recovery from simulated data*

132 To ensure that the models could recover known parameter values, we simulated ten data sets from

133 a bounded power law:

$$x_i = (\mathrm{u}_i x_{max}^{(\lambda+1)} + (1 - \mathrm{u}_i)x_{min}^{(\lambda+1)})^{\frac{(1}{(\lambda+1)}}$$

134 where x_i is the individual body size from the ith simulation, u_i is a unique draw from a

135 $Uniform(0, 1)$ distribution, and all other variables are the same as defined above. We set x_{min}

136 =1 and x_{max} = 1000, and simulated i = 1000 values from each of 10 λ's ranging from -2.2 to -1.2.

137 To generate $counts$, we rounded each simulated value to the nearest 0.001 and then tallied them.

138 We estimated the ten λ values in two ways. First, we fit a separate intercept-only model to each

139 of the ten data sets. Second, we fit a varying intercepts model (Gelman et al. 2014). The structure

140 of this model is $b = \alpha + \alpha_{[group]}$ where each group represents and offset from the mean value of

141 lambda.

142 Finally, we simulated data for a regression model with a single continuous predictor and a varying

143 intercept: $lambda = \alpha + \beta x + \alpha_{[group]}$, where α = -1.5, β = -0.1, and σ_{group} was 0.3 for n = 3

144 groups. The predictor x was continuous with 3 values (-2, 0, 2) and two replicates per value per

145 group for n = 18 λ's. After solving for λ from these known parameter values, we simulated 300

146 individuals from each λ using the procedure above, with x_{min} = 1 and x_{max} = 100. We fit this

147 model 20 times to measure variation in parameter recovery among model runs.

148 *Sample Size*

149 We examined sensitivity to sample size (number of individual body sizes) across three λ values (-2,

150 -1.6, -1.2). For each λ, we varied the number of simulated individuals from 2 to 2048, representing

151 a 2^n sequence with n ranging from 1 to 11. Each of the 11 densities was replicated 10 times

9

resulting in 110 datasets of individual body sizes. We fit each data set using separate intercept-only

paretocounts models and then plotted the resulting λ values as a function of sample size.

Case Studies

To examine model performance on empirical data, we re-ran a previously published analysis from of Edwards et al. (2020). In Edwards' study, size spectra exponents were first estimated separately for each sample using maximum likelihood. Then the modeled exponents were used as response variables in linear regression models. The goal was to test for linear changes in size spectra over three decades using bi-yearly size data of marine fishes collected from the International Benthic Trawl Survey (IBTS). The data set and original model results are available in the `sizeSpectra` package (Edwards et al. 2017). We tested the same hypothesis as Edwards et al. (2020), but instead of using a two-step process we fit a single model using the *paretocounts* lpdf (Eq. X).

Model Fitting

We fit each of the above models in rstan (Stan Development Team 2022) using 2 chains each with 1000 iterations. All models converged with r_{hat}'s <1.01. If a known parameter value fell inside the 95% Credible Intervals, we considered parameter recover successful. For the replicated regression model, we also tallied the number of times that the known value fell outside of the highest posterior density interval (HPI; out 20 runs).

Data Availability Statement

All data and code are available at https://github.com/jswesner/stan_spectra (to be permanently acrhived on acceptance).

Results

Parameter Recovery

For models fit to simulated individual data set, all credible intervals included the true value of λ and posterior medians were no more than 0.05 units away from the true value (Table 1). Similarly, when the same data set was fit using a varying intercepts model, the posterior median intercept α

Table 1: Table X. Parameter recovery.

Model	Parameter	True Value	q2.5	q50	q97.5
Separate Models	λ	-2.20	-2.22	-2.15	-2.09
Separate Models	λ	-2.09	-2.13	-2.06	-2.00
Separate Models	λ	-1.98	-2.05	-1.99	-1.92
Separate Models	λ	-1.87	-1.91	-1.85	-1.80
Separate Models	λ	-1.76	-1.83	-1.78	-1.72
Separate Models	λ	-1.64	-1.70	-1.66	-1.61
Separate Models	λ	-1.53	-1.59	-1.54	-1.50
Separate Models	λ	-1.42	-1.48	-1.44	-1.40
Separate Models	λ	-1.31	-1.31	-1.28	-1.24
Separate Models	λ	-1.20	-1.26	-1.23	-1.20
Single Model with Varying Intercepts	α	-1.70	-1.96	-1.71	-1.50
Single Model with Varying Intercepts	σ_[group]	0.34	0.23	0.36	0.58

177 and group standard deviation σ_{group} were nearly identical to the true values (Table 1).

178 Using the varying intercept model to estimate group specific means yielded similar results as using

179 separate models per group (Figure 1), demonstrating that a single model can be used to estimate

180 multiple size spectra simultaneously.

181 We also recovered regression parameters (intercept α and slope β) along with the group-level sd

182 (σ_{group}) (Figure 2). Fitting this model 20 times indicated reliable fit. For example, out of 60

183 comparison (3 parameters x 20 replicates) the true value fell outside of the 95% HDI only once

184 (Figure 2). Averaging the deviations (posterior median minus the true value) among the replicates

185 indicated no bias in the modeled estimates (mean bias +/- sd: $\alpha = 0$ +/- 0.05, $\beta = 0$ +/- 0.008, σ_{group}

186 = 0.03 +/- 0.06).

187 *Sample Size*

188 Variation in modeled estimates was high for samples containing less than 100 individual (Figure

189 2). For example, when the true λ value was -2, samples with just 8 individuals yielded estimates

190 ranging from -2.7 to -1.7. By contrast, all samples with more than 300 individuals captured the true

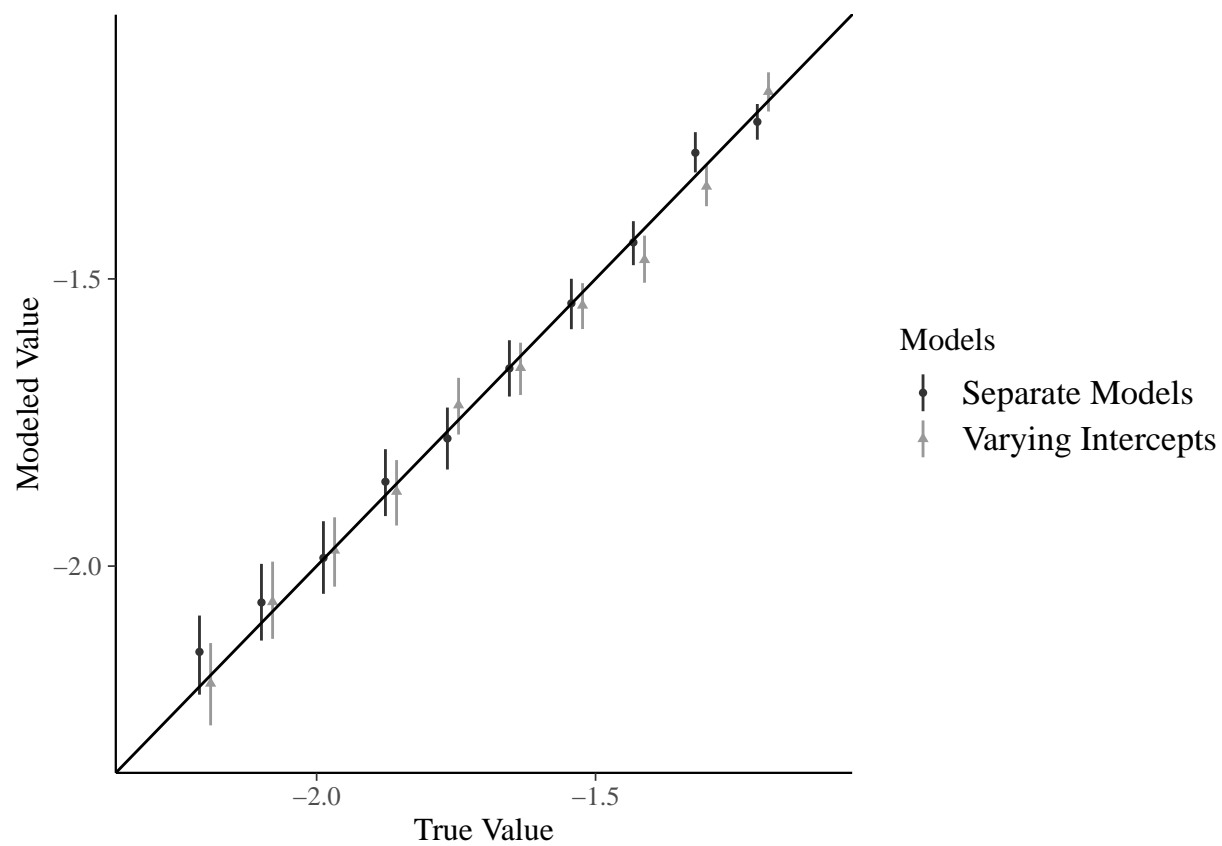191 λ with less than 0.1 unit of error (Figure 3).

192 *Case Study*

Figure 1: Modeled estimates (median +/- 95% Credible Intervals of λ using either 10 separate models or a single model with ten varying intercepts.
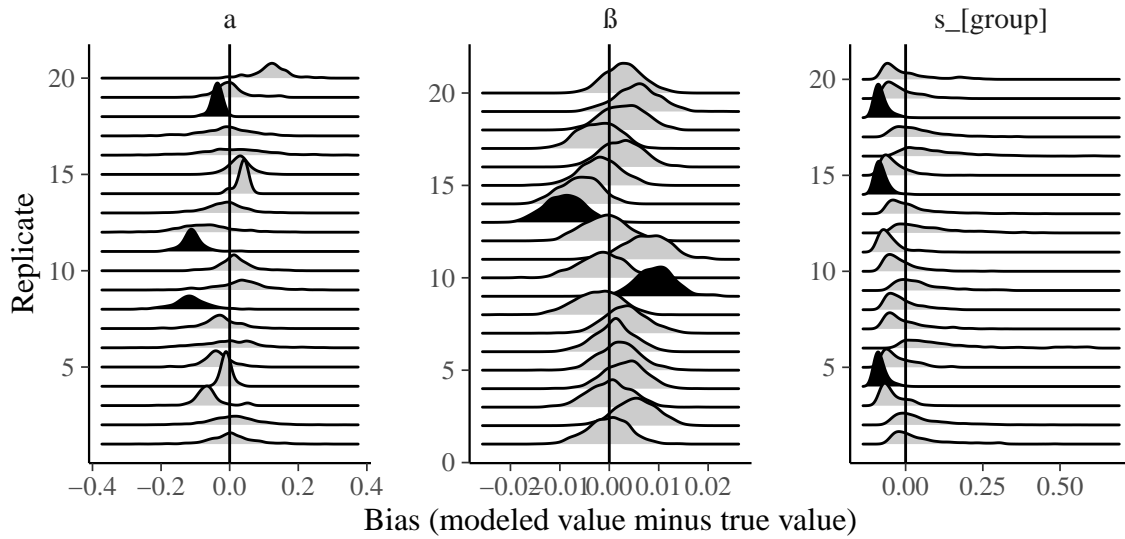
Figure 2: Posterior distributions of n = 20 modeled estimates of alpha, beta, and sigma_group for a linear regression estimating the size sepctrum exponent as a function of a continuous predictor. All data were simulated. Gray densities indicate that the HPI contains the true value, while black densities indicat the opposite. The vertical lines indicate true values.

Table 2: Table X. Slope values from a regression testing the relationship between the ISD exponent and year for IBTS trawl data (Edwards et al. 2020). The values are derived using the Bayesian hierarchical model presentd here or from the maximum likelihood approach described in Edwards et al. 2020).

Model	Mean	q2.5	q97.5
Bayesian - one step	-0.001	-0.005	0.002
MLE - two steps	-0.001	-0.005	0.003

Using IBTS data with the Bayesian hierarchical regression, we found a negative trend over time. The ISD exponent of IBTS trawl data declined by ~0.001 units per year, but with a 95% CrI ranging from -0.005 to 0.002. These values were nearly identical to those reported by Edwards et al. (2020) using a two-step approach (Table 3).

An advantage of fitting the model in a single Bayesian hierarchical framework is that estimates for individual groups are pulled toward the mean via partial pooling. This is apparent in comparing the unpooled MLE estimates (Figure Xa) to the partially pooled Bayesian estimates in each year (Figure Xb).
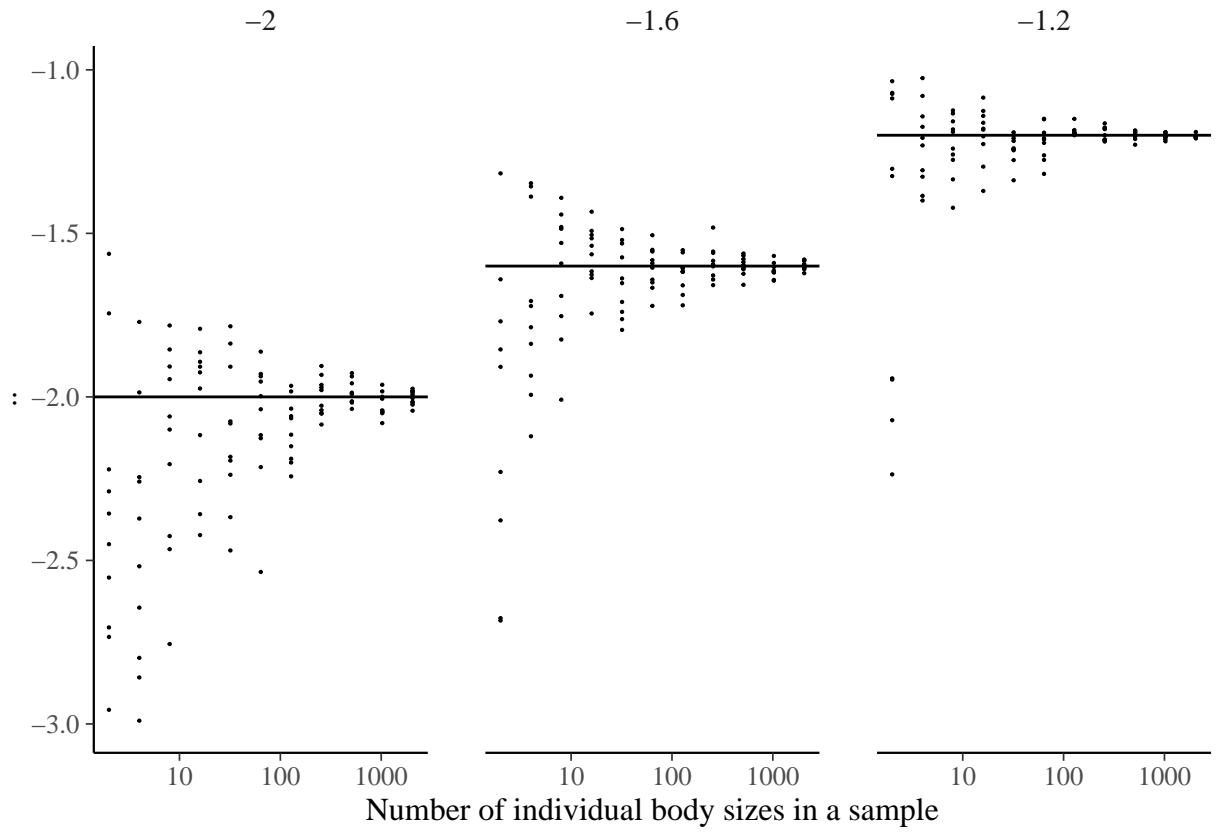
Figure 3: Estimates of λ across 11 different sample sizes (ranging from 2 to 2048 individuals) and three different true λ's (-2, -1.6, -1.2). Ten separate models were fit for each of the 11 sample sizes. The horizontal lines show the true value of λ
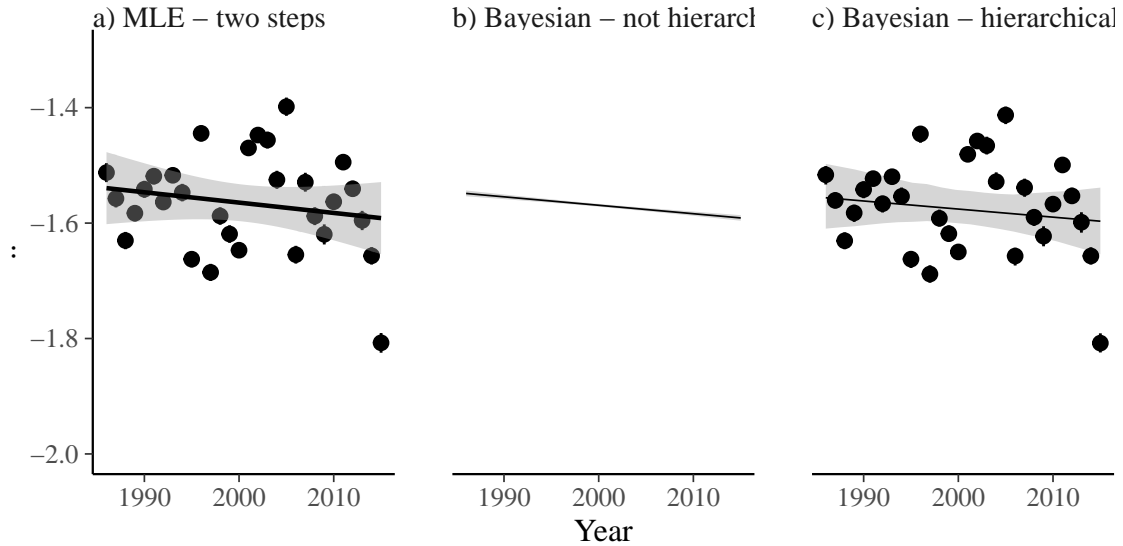
Figure 4: Regression results from a) Edwards et al. (2020) using maximum likelihood and linear regression (two steps), b) the Bayesian model with a paretocounts lpdf (one step), but without varying intercepts, and c) the Bayesian model with varying intercepts. In a) the points represent maximume likelihood estimates calculated separately for each year. In c) they represent hierarchical varying intercepts calculated from the model. There are no points in (b) because the model does not estimate lambda for individual years.

Discussion

The most important result of this work is the ability to analyze ISD exponents using fixed and random predictors in a hierarchical model. Our approach allows ecologists to test hypotheses about size spectra while avoiding the pitfalls of binning, which loses information and can lead to biased estimates of λ (White et al. 2008). Maximum likelihood solves this problem by directly estimating the ISD, but testing hypotheses with maximum likelihood still requires a two-step process in which λ is estimated individually for each sample and the results are then used as response variables in linear or non-linear models (Edwards et al. 2020). Our approach merges these steps, allowing for the incorporation of prior probabilities and hierarchical structure.

The ability to incorporate prior information using Bayesian updating has two practical advantages over analysis with maximum likelihood. First, adding informative prior distributions can improve model fit by limiting the MCMC sampler to reasonable sampling space. In other words it would not be sensible to estimate the probability that λ is -1,234 or -9. Without informative priors, those values

214 (and more extreme values) are considered equally likely and hence waste much of the algorithms

215 sampling effort on unlikely values (e.g., Wesner et al. 2021).

216 Second, and most importantly, ecologists have much prior information on the values that λ can

217 take. For example, Sheldon's conjecture suggests that λ is -2.05 (Andersen et al. 2006), a value

218 supported in pelagic marine ecosystems (Glazier et al. XXXX) and among forest trees (Enquist et

219 al. I THINK?). However, benthic marine systems typically have shallower exponents (e.g., ~-1.4;

220 Blanchard et al. 200X), while freshwater stream ecosystems have values somewhere in between

221 (-1.5; Pomeranz et al. 2022). However, despite apparent variation among ecocystems, there is a

222 strong indication that the values of λ are most likely to be somewhere between -2.05 and -1.4 for

223 many ecosystems. As a result, a prior that places most of its probability mass on these values

224 (e.g., $Normal(-1.6, 0.2)$ seems appropriate. Such a continuous prior does not prevent findings

225 of larger or smaller λ, but instead places properly weighted skepticism on such values.

226 Similar to priors, partial pooling from varying intercepts provides additional benefits, allowing for

227 the incorporation of hierarchical structure and pulling λ estimates towards the global mean (Gelman

228 et al. 2005, Qian et al. 2010). In the examples shown here, the amount of pooling is relatively

229 small because the sample sizes are large (>1000 individuals). However, the primary benefits of

230 pooling (both from varying effects and skeptical priors) is in prediction (Gelman et al. 2005, Hobbs

231 and Hooten 2015). This becomes especially important when models are used to forecast future

232 ecosystem conditions. We are unaware of efforts to forecast ISD's, but such forecasts seem to be

233 especially useful with modern long-term data sets like NEON (National Ecological Observatory

234 Network) in which body size samples will be collected at the continental scale over at least the next

235 20 years (Loescher et al. 2016). In addition, because the effects of priors and pooling increase with

236 smaller samples sizes, varying intercepts are likely to be particularly helpful for small samples. In

237 other words, priors and partial pooling contain built-in skepticism of extreme values, ensuring the

238 maxim that "extraordinary claims require extraordinary evidence".

239 One major drawback to the Bayesian modeling framework here is time. Bayesian models of even

minimal complexity must be estimated with Markov Chain Monte Carlo techniques. In this study, we used the Hamiltonion Monte Carlo (HMC) algorithm with a No-U-Turn Sampler (NUTS) via `rstan` (Stan Development Team 2022). Stan is substantially faster than other commonly used programs such as JAGS and WinBUGS, which rely on Gibbs sampling. For example, Stan is 10 to 1000 times more efficient than JAGS or WinBUGS, with the differences becoming greater as model complexity increases (Monnahan et al. 2017). In the current study, intercept-only models for individual samples with ~ 300 to 1500 individuals could be fit quickly (<2 seconds total run time (warm-up + sampling)) with as little as 1000 iterations and two chains. However, the IBTS regression models took >2 hours to run with the same iterations and chains. These times include the fact that our models used several optimization techniques, such as informative priors, standardized predictors, and non-centered parameterization, each of which are known to improve convergence and reduce sampling time (McElreath et al. 2020). But if Bayesian inference is desired, these runtimes may be worth the wait. In addition, they are certain to become even faster with the refinement of existing algorithms and the introduction of newer ones like Microcanonical HMC (Robnick et al. 2022).

Understanding how body sizes are distributed in ecological communities has been a focus in ecology for over a century (Peters 1983). Individual size distributions or community size spectra are an important focus of size-based research, and represent a fundamental measure of ecosystem structure and function (Kerr and Dickie 2001, White et al. 2007, Perkins et al. 2019). Most recently, ecologists have focused on testing theoretical predictions of how environmental changes affect ISD's (Nelson et al. 2017, O'Gorman et al. 2017, Perkins et al. 2019, Pomeranz et al. 2022). However, comprehensive analytical approaches to testing these hypotheses are lacking. We present a single analytical analytical approach that takes advantage of the underlying data structures of individual body sizes (Pareto distributions) while placing them in a generalized (Non)-linear hierarchical modeling framework. We hope that ecologists will adopt and improve on the models here to critically examine hypotheses of size spectra or other power-law distributed data.

Acknowledgements

17

References

TBD