Properties of Alternative Dirichlet-Multinomial Forms

for Fitting Composition Data

Jon Brodziak, NOAA Fisheries, Pacific Islands Fisheries Science Center

Jon.Brodziak@NOAA.GOV

17 July 2025

This document outlines the analytical framework for a study to evaluate the Dirichlet-multinomial distribution's effectiveness in fitting composition data within stock assessment models. Specifically, we examine two parameterizations of the Dirichlet-multinomial (DM) distribution, a linear form and a saturated form, and include the standard multinomial distribution for comparison. The multinomial model assumes independent trials with fixed probabilities for each category, which often underrepresents variability in real-world compositional data due to heterogeneity and clustering. In contrast, the DM approach models category probabilities as random draws from a Dirichlet distribution, capturing both overdispersion and inter-category correlation. The objective of the study is to assess which DM parameterization performs best across a range of stock assessment scenarios.

To facilitate this comparison, we define the notational framework and introduce two DM parameterizations as described by Thorson et al. (2017) and Fisch et al. (2021). The first step is to establish a glossary of key terms and symbols used throughout the study, which describe the data structure and parameter vectors relevant to fitting composition data (Table 1). This standardized terminology enables precise comparison and reproducibility across different likelihood formulations. Our intention is to provide a clear and consistent basis for the statistical modeling approaches employed.

Composition data are counts or proportions of individuals classified into predefined categories—such as age or size classes—capturing the distribution of characteristics within a statistical sampling frame. We will use age-structured data as the type of composition data sampled from the fish stock. The number of categories, or age classes, is denoted by K, with the age bins k=1, 2, …, (K-1) representing true ages and the and the age bin K representing all individuals K and older.

We use P_k to denote the proportion of individuals in category k. There are three types of proportion vectors used in this study. The true proportion vector at age in the stock is P^{TRUE}. The observed proportion vectors in the sampling frame are P^{OBS}. The predicted proportion vectors estimated by optimizing the parameters of the DM likelihood forms fitted to the observed data are P^{PRED}. These three proportion types are for simulating observed proportion data with a known sample size from the true population and estimating the predicted proportions. The simulated data will include observation error from random sampling and can also include process error due to random year-to-year fluctuations from environmental and demographic factors.

We use X_k to denote the count of individuals in category k. As with the proportion types, there are three types of count vectors used in this study. The true vector of absolute numbers at age in the stock is X^{TRUE}. The observed count vectors from the sampling frame are X^{OBS}. The predicted count vectors estimated by optimizing the parameters of the DM likelihood forms fitted to the observed data are X^{PRED}. Typically, the absolute and predicted counts of individuals are modeled as continuous variables for computational tractability while the observed composition data are discrete counts of individuals.

Table 1. Glossary of terms.

Variable	Description
K	Number of categories of age or size bins indexed by $k = 1, 2, \ldots, K$
$\mathbf{p}_t$	True population proportion vector at time t with elements $\mathrm{p}_{t,k}$
$\mathbf{p}_{obs,t}$	Observed proportion vector at time t with elements $\mathrm{p}_{obs,t,k}$
$\mathbf{p}_{pred,t}$	Predicted proportion vector at time t with elements $\mathrm{p}_{pred,t,k}$
$\mathbf{N}$	Population numbers at time t in category k, with elements $\mathrm{N}_{t,k}$
$\mathbf{x}_{obs,t}$	Observed count vector at time t with elements $\mathrm{x}_{obs,t,k}$
$\mathbf{x}_{pred,t}$	Predicted count vector at time t with elements $\mathrm{x}_{pred,t,k}$
n_t	Observed composition sample size at time t
$\boldsymbol{\alpha}$	Concentration parameter vector for the Dirichlet-multinomial distribution
α_0	Sum of concentration parameter vector for the Dirichlet-multinomial distribution
θ	Weighting parameter for the linear Dirichlet-multinomial distribution

β	Weighting parameter for the saturating Dirichlet-multinomial distribution
$n_{eff,t}$	Effective sample size for the observed proportions at time *t*

The multinomial negative loglikelihood NLL_{MN} for composition data in a single time period *t* is expressed as

$$(1.1) \qquad NLL_{MN}\left(\mathbf{p}_{pred,t} \,|\, n_t,\, \mathbf{p}_{obs,t}\right) = -n_t \sum_{c} \mathrm{p}_{obs,t,k} \cdot \ln\left(\mathrm{p}_{pred,t,k}\right) - \ln\left(n_t\,!\right) + \sum_{k=1}^{K} \ln\left(\mathrm{x}_{obs,t,k}\,!\right)$$

where boldface indicates a *K*-dimensional vector.

In equation (1.1) the parameters of the NLL_{MN} are in the vector of predicted proportions $\mathbf{p}_{pred,t} = \left(\mathrm{p}_{pred,t,1}, \ldots, \mathrm{p}_{pred,t,K}\right)$ at time *t* as a function of the input sample size n and the observed proportions $\mathbf{p}_{obs,t} = \left(\mathrm{p}_{obs,t,1}, \ldots, \mathrm{p}_{obs,t,K}\right)$.

The multinomial negative loglikelihood NLL_{MN} for composition data in a set of *T* time periods indexed by *t* is expressed as

$$(1.2) \qquad NLL_{MN}\left(\mathbf{P}_{pred,t} \,|\, \mathbf{n}, \mathbf{P}_{obs,t}\right) = -\sum_{t=1}^{T}\left[n_t \sum_{c} \mathrm{p}_{obs,t,k} \cdot \ln\left(\mathrm{p}_{pred,t,k}\right) - \ln\left(n_t\,!\right) + \sum_{k=1}^{K} \ln\left(\mathrm{x}_{obs,t,k}\,!\right)\right]$$

where capitalized boldface indicates a *K*x*T*-dimensional matrix.

In equation (1.2) the parameters of the NLL_{MN} are in the matrix of predicted proportions $\mathbf{P}_{pred,t} = \left[\mathrm{p}_{pred,t,k}\right]_{KxT}$ at time *t* as a function of the input sample size vector $\mathbf{n}$ and the matrix of observed proportions $\mathbf{P}_{obs,t} = \left[\mathrm{p}_{obs,t,k}\right]_{KxT}$.

The linear Dirichlet-multinomial negative loglikelihood NLL_{DML} for composition data in a single time period *t* is expressed as

$$\begin{aligned} NLL_{DML}\left(\mathbf{p}_{pred,t},\theta \mid n_t,\, \mathbf{p}_{obs,t}\right) = & -\ln\left(\Gamma\left(n_t+1\right)\right)+\sum_{k=1}^{K}\ln\left(\Gamma\left(n_t\cdot \mathrm{p}_{obs,t,k}+1\right)\right) \\ & -\ln\left(\Gamma\left(\theta n_t\right)\right)+\ln\left(\Gamma\left(n_t+\theta n_t\right)\right) \\ & -\sum_{k=1}^{K}\ln\left(\Gamma\left(n_t\cdot \mathrm{p}_{obs,t,k}+\theta n_t\cdot \mathrm{p}_{pred,t,k}\right)\right)+\sum_{k=1}^{K}\ln\left(\Gamma\left(\theta n_t\cdot \mathrm{p}_{pred,t,k}\right)\right) \end{aligned} \tag{1.3}$$

where θ is the linear Dirichlet-multinomial weighting parameter. In equation (1.3) the parameters of the NLL_{DML} are in the vector of predicted proportions $\mathbf{p}_{pred,t}$ at time *t* and the weighting parameter θ as a function of the input sample size n_t and the observed proportions $\mathbf{p}_{obs,t}$.

The linear Dirichlet-multinomial negative loglikelihood for a set of *T* time periods indexed by *t* is expressed as

(1.4)

$$NLL_{DML}\left(\mathbf{P}_{pred,t},\theta \middle| \mathbf{n},\mathbf{P}_{obs,t}\right)=\sum_{t=1}^{T}\left\{\begin{aligned} & -\ln\left(\Gamma\left(n_t+1\right)\right)+\sum_{k=1}^{K}\ln\left(\Gamma\left(n_t\cdot \mathrm{p}_{obs,t,k}+1\right)\right) \\ & -\ln\left(\Gamma\left(\theta n_t\right)\right)+\ln\left(\Gamma\left(n_t+\theta n_t\right)\right) \\ & -\sum_{k=1}^{K}\ln\left(\Gamma\left(n_t\cdot \mathrm{p}_{obs,t,k}+\theta n_t\cdot \mathrm{p}_{pred,t,k}\right)\right)+\sum_{k=1}^{K}\ln\left(\Gamma\left(\theta n_t\cdot \mathrm{p}_{pred,t,k}\right)\right) \end{aligned}\right\}$$

In equation (1.4) the parameters of the NLL_{DML} are in the matrix of predicted proportions $\mathbf{P}_{pred,t}$ at time *t* and the weighting parameter θ as a function of the input sample size vector $\mathbf{n}$ and the matrix of observed proportions $\mathbf{P}_{obs,t}$.

The saturated Dirichlet-multinomial negative loglikelihood NLL_{DML} for composition data in a single time period *t* is expressed as

$$\begin{aligned} NLL_{DMS}\left(\mathbf{p}_{pred,t},\beta \mid n_t,\, \mathbf{p}_{obs,t}\right) = & -\ln\left(\Gamma\left(n_t+1\right)\right)+\sum_{k=1}^{K}\ln\left(\Gamma\left(n_t\cdot \mathrm{p}_{obs,t,k}+1\right)\right) \\ & -\ln\left(\Gamma\left(\beta\right)\right)+\ln\left(\Gamma\left(n_t+\beta\right)\right) \\ & -\sum_{k=1}^{K}\ln\left(\Gamma\left(n_t\cdot \mathrm{p}_{obs,t,k}+\beta\cdot \mathrm{p}_{pred,t,k}\right)\right)+\sum_{k=1}^{K}\ln\left(\Gamma\left(\beta\cdot \mathrm{p}_{pred,t,k}\right)\right) \end{aligned} \tag{1.5}$$

where β is the saturated Dirichlet-multinomial weighting parameter. In equation (1.5) the parameters of the NLL_{DMS} are in the vector of predicted proportions $\mathbf{p}_{pred,t}$ at time *t* and the weighting parameter β as a function of the input sample size n_t and the observed proportions $\mathbf{p}_{obs,t}$.

The saturated Dirichlet-multinomial negative loglikelihood for a set of *T* time periods indexed by *t* is expressed as

(1.6)

$$NLL_{DMS}\left(\mathbf{P}_{pred,t},\beta|\mathbf{n},\mathbf{P}_{obs,t}\right)=\sum_{t=1}^{T}\left\{\begin{array}{l}-\ln\left(\Gamma\left(n_t+1\right)\right)+\sum_{k=1}^{K}\ln\left(\Gamma\left(n_t\cdot \mathrm{p}_{obs,t,k}+1\right)\right)\\ -\ln\left(\Gamma\left(\beta\right)\right)+\ln\left(\Gamma\left(n_t+\beta\right)\right)\\ -\sum_{k=1}^{K}\ln\left(\Gamma\left(n_t\cdot \mathrm{p}_{obs,t,k}+\beta\cdot \mathrm{p}_{pred,t,k}\right)\right)+\sum_{k=1}^{K}\ln\left(\Gamma\left(\beta\cdot \mathrm{p}_{pred,t,k}\right)\right)\end{array}\right\}$$

In equation (1.6) the parameters of the NLL_{DMS} are in the matrix of predicted proportions $\mathbf{P}_{pred,t}$ at time *t* and the weighting parameter β as a function of the input sample size vector $\mathbf{n}$ and the matrix of observed proportions $\mathbf{P}_{obs,t}$.

The effective sample size $n_{eff,t}$ for the linear Dirichlet Multinomial in period *t* depends on θ and n_t as

(1.7)

$$n_{eff,t}=\frac{1+\theta n_t}{1+\theta}$$

And for a set of *T* periods n_{eff} is

(1.8)

$$n_{eff}=\sum_{t=1}^{T}n_{eff,t}=\sum_{t=1}^{T}\frac{1+\theta n_t}{1+\theta}$$

Similarly, the effective sample size $n_{eff,t}$ for the saturated Dirichlet Multinomial in period *t* depends on β and n_t as

(1.9)

$$n_{eff,t}=\frac{n_t+\beta n_t}{n_t+\beta}$$

And for a set of periods n_{eff} is

(1.10)

$$n_{eff}=\sum_{t=1}^{T}n_{eff,t}=\sum_{t=1}^{T}\frac{n_t+\beta n_t}{n_t+\beta}$$

First order partial derivatives for the effective sample size $n_{eff,t}$ of the linear Dirichlet Multinomial as a function of the input parameters are:

(1.11) $$\frac{\partial n_{eff,t}}{\partial n_t} = \frac{\theta}{1+\theta} = \frac{\theta}{1+\theta n_t} n_{eff,t}$$

And

(1.12) $$\frac{\partial n_{eff,t}}{\partial \theta} = \frac{n_t - 1}{\left(1+\theta\right)^2} = \frac{n_t - 1}{\left(1+\theta\right)\left(1+\theta n_t\right)} n_{eff,t}$$

Second order partial derivatives for the effective sample size n_{eff} of the linear Dirichlet Multinomial as a function of the input parameters are:

(1.13) $$\frac{\partial^2 n_{eff,t}}{\partial n_t^2} = 0$$

(1.14) $$\frac{\partial^2 n_{eff,t}}{\partial n \partial \theta} = \frac{1}{\left(1+\theta\right)^2}$$

(1.15) $$\frac{\partial^2 n_{eff,t}}{\partial \theta_t^2} = \frac{-2(n_t - 1)}{\left(1+\theta\right)^3}$$

Similarly, first order partial derivatives for the effective sample size $n_{eff,t}$ of the saturated Dirichlet Multinomial as a function of the input parameters are:

(1.16) $$\frac{\partial n_{eff,t}}{\partial n_t} = \frac{\beta\left(1+\beta\right)}{\left(n_t+\beta\right)^2} = \frac{\beta}{n_t\left(n_t+\beta\right)} n_{eff,t}$$

And

(1.17) $$\frac{\partial n_{eff,t}}{\partial \beta} = \frac{n_t\left(n_t - 1\right)}{\left(n_t+\beta\right)^2} = \frac{\left(n_t - 1\right)}{\left(1+\beta\right)\left(n_t+\beta\right)} n_{eff,t}$$

Second order partial derivatives for the effective sample size n_{eff} of the saturated Dirichlet Multinomial as a function of the input parameters are:

(1.18) $$\frac{\partial n_{eff,t}^2}{\partial^2 n_t} = \frac{-2\left(\beta^2+\beta\right)}{\left(n_t+\beta\right)^3}$$

$$\frac{\partial n_{eff,t}^{2}}{\partial \beta^{2}} = \frac{-2n^{2} + 2n}{\left(n_{t} + \beta\right)^{3}} \tag{1.19}$$

References

Fisch, N., Camp, E., Shertzer, K., Ahrens, R. 2021. Assessing likelihoods for fitting composition data within stock assessments, with emphasis on different degrees of process and observation error. Fisheries Research, 243, 106069, https://doi.org/10.1016/j.fishres.2021.106069

Thorson, J.T., Johnson, K.F., Methot, R.D., Taylor, I.G. 2017. Model-based estimates of effective sample size in stock assessment models using the Dirichlet-multinomial distribution. Fisheries Research, 192:84-93, https://doi.org/10.1016/j.fishres.2016.06.005