A Novel Method to Estimate Prey Contributions to Predator Diets

<table>
<thead>
<tr>
<th>Journal:</th>
<th>Canadian Journal of Fisheries and Aquatic Sciences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>cjfas-2016-0149.R3</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Article</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>08-Jun-2016</td>
</tr>
</tbody>
</table>
| Complete List of Authors: | Moriarty, Pamela; University of Washington, School of Aquatic and Fishery Sciences  
Essington, Timothy; University of Washington, School of Aquatic and Fishery Sciences  
Ward, Eric; National Marine Fisheries Service - NOAA, Northwest Fisheries Science Center |
| Keyword: | Mixture Model, Diet Estimate, Food Web, Predator-Prey |
A Novel Method to Estimate Prey Contributions to Predator Diets

Pamela E. Moriarty¹*, Timothy E. Essington², Eric J. Ward³

¹*corresponding author

pmoriart@uw.edu

240-505-0693

School of Aquatic and Fishery Sciences
University of Washington
Box 355020
Seattle, WA 98195

2- essing@uw.edu

School of Aquatic and Fishery Sciences
University of Washington
Box 355020
Seattle, WA 98195

3- eric.ward@noaa.gov

Northwest Fisheries Science Center, NOAA
2725 Montlake Boulevard
Seattle, WA 98112

Code for the model is available at https://github.com/pmoriart/DietMixtureModel
ABSTRACT

Stomach content data are frequently used to characterize predator feeding habits, often by describing the proportional contribution by weight or number of each prey type (diet fractions). These data pose several statistical challenges for analysis and estimation that have hindered our ability to create quantitative diet fraction estimates from stomach content data. To address these challenges, we developed a novel, likelihood-based mixture model to quantitatively estimate diet fractions. Simulation testing indicated that estimated diet fractions from the mixture model were more precise than those estimated either from a (stomach-mass) weighted mean or the sample mean, and were more accurate than a sample mean. Additionally, we applied the mixture model, a weighted mean, and sample mean to stomach content data for multiple types of predators. For three of four of these datasets, the mixture model demonstrated higher precision than and similar accuracy to a weighted-mean, and similar precision and better accuracy than a sample mean. The mixture model represents an important step in advancing statistical methods to address the challenges of stomach content data.
Introduction

Ecological research fundamentally requires understanding the feeding habits of predators. From a behavioral ecology perspective, we are commonly interested in measuring prey breadth, and the factors influencing niche widths (Catano et al. 2014; Charnov 1976; Navia et al. 2007; Randall and Myers 2001). Data on predator feeding habits are also used to better understand top-down and bottom-up effects in predator-prey systems, including how distinct patterns of prey use might govern predator-prey interaction strengths (Marshall 2007; Polis and Hurd 1996). These data are also essential for creating food webs and understanding their structures and dynamics (Huxel and McCann 1998; McCann et al. 1998; Prugh 2005). Finally, feeding habits information is critical for parameterizing models to predict the outcome of predator–prey interactions (Christensen and Walters 2004; Fulton et al. 2011; Rose et al. 2008; Stewart et al. 1981).

Despite the pervasive role of feeding habit information, there has been surprisingly little development of methods for quantitative estimation of predator diets.

For many species, predator feeding information is often derived from sampling stomach contents (i.e. stomach content data). Stomach content analysis is ubiquitously used in studies of fish diets (MacKinlay and Shearer 1996), as well as for seabirds (Montalti and Ruben Coria 1993), jellyfish (Barz and Hirche 2007; Jaspers et al. 2015), sea turtles (Colman et al. 2014; Williams et al. 2014), and marine mammals (Fernandez et al. 2014; Matley et al. 2015). Analysis of stomach content data collected involves two distinct components: 1) measuring total prey amounts in the stomach (total stomach contents’ mass or number) (Elliott and Persson 1978) and 2) calculating the fraction of total prey that consists of individual prey types (diet fractions). While there is a large
body of literature examining distinct ways to analyze stomach content data, there has
been little development of formal statistical methods.

There are few commonly used mass-based analysis methods for estimating diet
fractions. One is to use the average of individual diet fractions, treating each stomach as
an independent data point (Liao et al. 2001; Nielsen et al. 1983). Alternatively, all
stomachs can be pooled and the proportion of the total summed diet mass consisting of
each prey type can be calculated (Nielsen et al. 1983). Additionally, a weighted mean
(weighting individual stomach proportions by total stomach contents’ mass) is sometimes
employed, which is analogous to the pooling method, but allows for calculation of
confidence intervals (Nielsen et al. 1983).

These analyses are relatively straightforward to compute, but unfortunately don’t
address the challenging characteristics of stomach content data. One challenge of
analyzing stomach content data is that feeding rates may be related to the prey type
consumed- when certain prey items are present they may be likely to be consumed in
high quantities. For example, on the occasions that rarely available schooling prey is
available, a generalist predator may consume it at high rates. This creates covariance at
the individual level between the amount a predator consumes and the prey type eaten that
is not reflective of population-level consumption (Hyslop 1980; Liao et al. 2001).

Another challenge is that stomach content data are often skewed such that a small number
of samples have much higher stomach content masses than the average of all the samples.
At a population level, these rare events could comprise a high proportion of total feeding,
so are important data to include in predator diet estimates. However, rare events are
difficult to estimate precisely, and their presence in a small sample can greatly influence
diet fraction estimates (Ahlbeck et al. 2012).

Methods for stomach content data should ideally be based on an appropriate likelihood function to estimate the prey contribution. Likelihood functions for individual sample proportions are possible, but are generally not existent for weighted mean proportions. However, the use of a likelihood based framework in ecology and fisheries, whether in maximum likelihood (ML) or Bayesian models, offers several advantages (Hilborn and Mangel 1997). Likelihood based estimation of parameters allows for formal model selection, using criteria such as AIC (Burnham and Anderson 1998), so that distinct alternative hypotheses can be evaluated in a statistically rigorous way. As diet estimates are used as inputs for other models (Christensen and Walters 2004; Fulton et al. 2004; Plagányi et al. 2014), formally selecting the best estimate could allow for more accurate inputs to other models. A likelihood framework also allows for merging prior information on predator diets (such as stable isotope analyses, expert judgment or previous sampling) with stomach content data (Ainsworth et al. 2010). Finally, likelihood based approaches can be extended to address the potential lack of independence among stomachs collected during a single sampling event.

Here, we present a novel statistical method to quantitatively estimate prey contributions to a predator’s diet that addresses all three challenges discussed above: 1) extreme events in predator stomach contents’ mass (defined here as an individual stomach having a total contents mass greater than twice the interquartile range), 2) covariance between stomach contents’ mass and diet fraction within the data and 3) uses a maximum-likelihood framework. Then, we demonstrate the use of the mixture model by applying it to datasets for different predators with varying life histories.
METHODS

112 We begin by explaining the construction of our model to quantitatively estimate predator diet fractions. Then we evaluate the sensitivity of our model to simplifying assumptions using simulated data and compare its accuracy and precision to existing methods for analyzing stomach content data. Finally, we apply our model to datasets for different predators to determine how the methods compare in real world situations.

119 Model Development

Diet fraction, defined as the mean proportional contribution of a single prey type, \( i \), to a predator’s diet, \( C_i \) (Table 1), is equal to the ratio of the expected mass of prey type \( i \) in the stomachs, \( E[M_i] \), divided by the expected total mass of the stomachs’ contents, \( E[M_s] \),

\[
E[C_i] = \frac{E[M_i]}{E[M_s]},
\]

The desired properties of an estimator for \( E[C_i] \) are that it, and its confidence intervals, are bound between 0 and 1, inclusively, and that it is based on a likelihood function to permit model selection and maximum likelihood estimation. There are two existing distinct frameworks, the mean and weighted mean, to estimate \( E[C_i] \). Using the mean proportion of prey type \( i \) in individual stomachs implies that given stomachs from \( N \) individual consumers, each with prey content mass \( M_{ij} \) and total stomach content mass \( M_{sj} \), where \( 1 < j < N \), \( E[C_i] = \text{mean}(p_i) = \frac{\sum_{j=1}^{N} M_{ij}}{M_{sj}} \). In contrast, using the weighted mean or pooling all stomachs is equivalent to estimating \( E[C_i] = \text{weighted mean} \ (p_i, M_i) = \)
\[
\sum_{i=1}^{N} \frac{M_{ij}}{\sum_{j=1}^{N} M_{sj}}. \text{ If } M_{ij} \text{ and } M_{sj} \text{ are independent, these two sample-based methods will, on average, produce equivalent estimates.}
\]

**Likelihood Estimation**

To create a likelihood based framework that will ensure the estimator is bound between 0 and 1, we model \( M_i \) as the product of the total stomach contents’ mass \( M_s \) and the mean fraction of diet that consists of prey type \( i (P_i) \), where both stomach mass and stomach diet fraction are now random variables, described by probability density functions \( f(M_s) \) and \( g(P_i) \), respectively. However, to account for the fact that \( P_i \) and \( M_s \) may not be independent, we allow \( g(P_i) \) to also be a function of stomach content mass, \( g(P_i, M_s) \), so the mean of \( g(P_i) \) depends on \( M_s \). Thus, in our framework, the numerator of equation 1 is the integration over \( P_i \) and \( M_s \):

\[
\int_0^1 \int_{\mathbb{R}^+} g(P_i, M_s) P_i M_s f(M_s) dM_s dP_i.
\]

The denominator is obtained by integrating over \( M_s \) in the typical way, producing:

\[
E[C_i] = \frac{\int_0^1 \int_{\mathbb{R}^+} g(P_i, M_s) P_i M_s f(M_s) dM_s dP_i}{\int_{\mathbb{R}^+} M_s f(M_s) dM_s},
\]

In the simple case where \( g(P_i, M_s) \) is independent of \( M_s \), this simplifies to

\[
\int_0^1 P_i g(P_i) dP_i,
\]

which is simply the expectation of \( P_i \). When \( g(P_i, M_s) \) depends on \( M_s \), the integral in the numerator of equation 2 has no analytical solution except when \( g(P_i, M_s) \) is linear in \( M_s \). However, this case is generally not practical because \( P_i \) must be bound between 0 and 1, and typical methods to achieve this constraint, such as piecewise linear models, would not produce an analytical solution.

Additionally, there is a second calculation problem when using equation 2. This equation requires a probability density function that allows for a greater than zero density when \( P_i \) equals 0 or 1, but has zero density when \( P_i \) is less than 0 or greater than 1. No
continuous probability density function has this property. One solution for a likelihood
function is to create a mixture model and introduce two new variables. The first, $\Phi$, is an
indicator variable to represent the presence of prey type $i$ in a stomach and is described
by $r_\Phi$, the probability of presence. Correspondingly, stomachs not containing the prey
type $i$ are denoted by $\neg \Phi$. The second random variable, $\Phi_{p=1}$, indicates the stomach
consisted entirely of prey type $i$ given it contained prey type $i$ with the probability of this
event denoted by $r_{\Phi_{p=1}}$.

Although including mixture models at first appears to complicate the analytical
formulation, it provides an opportunity to account for potential covariance between $P_i$
and $M_s$. Namely, we make two assumptions regarding how stomach mass is related to
stomach contents. One, we assume that the probability density function for $M_s$ differs
between stomachs that contain prey type $i$ and those that do not. Two, we assume there is
no additional covariance between $P_i$ and $M_s$ among stomachs that contain the prey (Figure
1). Under these assumptions, the model simplifies greatly, such that the numerator of
equation 2 becomes:

$$E[M_i] = r_\Phi (1 - r_{\Phi_{p=1}}) E(P_i | \Phi, 1 - \Phi_{p=1}) + r_\Phi r_{\Phi_{p=1}} E(M_s | \Phi)$$

And the denominator of equation 2 can be expressed as:

$$E[M_s] = r_\Phi E(M_s | \Phi) + (1 - r_\Phi) E(M_s | 1 - \Phi)$$

Then, the full expectation is

$$E[C_i] = \frac{r_\Phi (1 - r_{\Phi_{p=1}}) E(P_i | \Phi, 1 - \Phi_{p=1}) + r_\Phi r_{\Phi_{p=1}} E(M_s | \Phi)}{r_\Phi E(M_s | \Phi) + (1 - r_\Phi) E(M_s | 1 - \Phi)}.$$

Another advantage of this mixture formulation is it consists of distinct parameters that
can be estimated separately from the data using maximum likelihood estimation (MLE).

Here, the probability of prey type $i$ occurring is estimated using a Bernoulli probability
density function, \( \Phi \sim Ber(\theta_\Phi) \), where \( \Phi \) indicates the presence of prey type \( i \). Similarly, the probability of only prey type \( i \) occurring given that prey type \( i \) has occurred is also estimated from a Bernoulli density function, \( \Phi_{p_i=1} \sim Ber(\theta_{\Phi_{p_i=1}}) \), where \( \Phi_{p_i=1} \) indicates only the presence of prey type \( i \) given prey type \( i \) has occurred. A common choice for estimating proportional data is a beta distribution, so we model \( g(P_i) \) as a beta distribution, \( P_i \sim Beta(\alpha_1, \alpha_2) \). The mean diet fraction can be written as the mean of the beta distribution with parameters \( \alpha_1 \) and \( \alpha_2 \), \( \alpha_1/(\alpha_1 + \alpha_2) \). The probability densities of the total stomach contents masses for those containing prey type \( i \), \( f(M_s|\Phi) \), and those not containing prey type \( i \), \( f(M_s|\neg\Phi) \), could be modeled using any non-negative continuous distribution. Because of its flexibility, we used a gamma distribution, so \( M_s|\Phi \sim Gamma(k_\Phi, 1/\theta_\Phi) \), with \( E[M_s|\Phi] = k_\Phi \theta_\Phi \) and \( M_s|\neg\Phi \sim Gamma(k_\alpha, 1/\theta_\alpha) \), so \( E[M_s|\neg\Phi] = k_\alpha \theta_\alpha \). The gamma distribution’s thick tails, as compared to other common choices (normal, lognormal), allow us to model the stomach data while including the stomachs with extreme content masses in the distribution’s long tails.

Using these probability density functions, the total likelihood function for the model we present here is expressed as three components: diet fractions equal to 0, diet fractions between 0 and 1, and diet fractions for single-prey stomachs:

\[
L = \prod_j \left\{ \begin{array}{l}
\frac{1}{\Gamma(k_\Phi)\theta_\Phi^{k_\Phi}} p_{ij}^{k_\Phi-1} e^{-\frac{p_{ij}}{\theta_\Phi}} * (1 - \theta_\Phi) \quad p_{ij} = 0 \\
\frac{1}{\Gamma(k_\alpha)\theta_\alpha^{k_\alpha}} p_{ij}^{k_\alpha-1} e^{-\frac{p_{ij}}{\theta_\alpha}} * \theta_\Phi * r_{\Phi p=1} * \frac{p_{ij}^{\alpha_1-1}(1-p_{ij})^{\alpha_2-1}}{\text{Beta}(\alpha_1, \alpha_2)} \quad 0 < p_{ij} < 1 \quad . \\
\frac{1}{\Gamma(k_\alpha)\theta_\alpha^{k_\alpha}} p_{ij}^{k_\alpha-1} e^{-\frac{p_{ij}}{\theta_\alpha}} * \theta_\Phi * r_{\Phi p=1} \quad p_{ij} = 1
\end{array} \right.
\]

[6]
(Note that we follow standard statistical notation by using upper case variables to indicate random variables and lower case letters to indicate outcomes, i.e. data.)

**Parameterization**

For the purpose of parameter estimation and calculating error, we reparameterized the model so that the main parameter of interest, $E[C_i] = \hat{\psi}_i$, was a leading parameter. That is, rather than estimate the eight components of the model independently, and use those to generate $\hat{\psi}_i$ as a derived parameter, we rearranged the model form so that $\hat{\psi}_i$ was being estimated simultaneously with the remaining parameters. This allowed us to estimate the variance – covariance matrix of $\hat{\psi}_i$ and other parameters of interest. Asymptotic error for the parameters was calculated using the inverted hessian matrix. Specifically, we rearranged the equation such that $M_s|\Phi$ was calculated as a derived parameter,

$$E[M_s|\Phi] = \frac{E[M_s|\Phi]r_\phi(E[P_i]+r_\phi r_{\Phi P=1}-E[C_i]-E[P_i]r_\phi r_{\Phi P=1})}{E[C_i](1-r_\phi)}.$$ 

Additionally, the gamma and beta likelihood functions each require estimation of two parameters that together give the mean and variance of the distribution. To aid model interpretation, we directly estimated the mean and variance for these distributions (i.e., $E[P_1], Var[P_1]$), rather than the distribution parameters (i.e., $\alpha_1, \alpha_2$), as the mean and variance have clear biological meanings in this context.

**Model Testing**

We tested the mixture model via simulation analysis. Specifically, we generated data with particular properties and a known diet fraction from an operating model (with a
known “truth”), fit our estimation model to these data, and then compared the true and estimated values. We performed several simulation tests; each asked specific questions about the performance of the model under particular conditions. Results were compared across three modeling approaches: (1) the mixture model, (2) the conventional sample mean, (3) the conventional sample weighted mean.

**Base Operating Model**

We generated data so that each stomach had a specified probability, $r_{\Phi}$, of having $p_{ij}>0$. For stomachs with $p_{ij}>0$, we then assigned $p_{ij}=1$ with probability $r_{\Phi_{p}=1}$. Stomach masses were drawn for $p_{ij}=0$ from a gamma distribution, $m_{sj}|\neg\Phi \sim \text{Gamma}(k_{a}, \theta_{a})$. For stomachs for which $p_{ij}$ was assigned to be $>0$, then the bivariate data ($p_{ij}$, $m_{sj}$) for each stomach was calculated from

$$\begin{align*}
(p_{ij}, m_{sj}) = \begin{cases}
  p_{ij} \sim \beta(\alpha_{1}, \alpha_{2}), m_{sj} \sim \text{Gamma}(k_{\Phi}, \theta_{\Phi}), & \text{if } p_{ij} \neq 1 \\
  p_{ij} \sim 1, m_{sj} \sim \text{Gamma}(k_{\Phi}, \theta_{\Phi}), & \text{if } p_{ij} = 1
\end{cases}
\end{align*}$$

For each model testing scenario, we compared the mean and variance of the prey contribution, $\bar{c}_{i}$, estimated by the mixture model to those estimated using the conventional mean and conventional weighted mean, as well as to the true value from the operating model. This allowed us to compare the accuracy and precision of the commonly used methods to the new method. We generated 100 data sets, each containing 200 stomachs, for each parameter value tested. Preliminary testing indicated that generating and estimating more than 100 datasets did not alter results or conclusions. Also, preliminary testing on the effect of sample size indicated that sample size affected model precision, but not accuracy. All parameter estimation was performed in R3.1.0 (R Core Team 2014).
To confirm that the numerical routines were converging on the parameter values that maximize the likelihood (MLEs), we also calculated the MLE for each parameter individually in addition to estimating them jointly. The estimate for $\hat{c}_i$ was then calculated from these individual estimates using equation 5. If the estimate of $\hat{c}_i$ from estimating all parameters jointly was within 20% of the MLE of $\hat{c}_i$ obtained by estimating them individually, then we considered the estimation to have converged on the MLE. This confirmation procedure was used throughout all model testing analyses.

**Mean Absolute Error**

Our first test was designed to determine whether the model parameters could be reliably estimated when all of the model assumptions were true. For this, we generated data directly from the base operating model and estimated the parameters with the model we developed. We calculated the mean absolute error (MAE) for each parameter in the model as,

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |\text{estimated} - \text{true}|.
\]

**Overdispersion**

Our second test was to evaluate whether the model estimates were robust to overdispersion in stomach contents mass, the distribution of diet fractions, or the presence / absence of prey in stomachs. To simulate overdispersion, we generated data from the basic operating model described above, but treated the distribution parameters of $\Phi$, $\Phi_{p=1}$ and $p_i$ as random. For $\Phi$ and $\Phi_{p=1}$, $R_\Phi$ and $R_{\Phi_{p=1}}$ were drawn from a beta distribution, so they would be bounded between 0 and 1, while the parameters for $P_i$, $A_i$
and $A_2$, were drawn from a gamma distribution, so they would be greater than 0. Other parameter values were held constant between simulations (Table 2).

Extreme Events in Total Stomach Contents’ Mass

We evaluated the model’s accuracy in estimating $\bar{c}_i$ under scenarios in which the data contained extreme events in stomach contents’ masses (e.g. a few very full stomachs). First, we used the basic operating model, but each stomach had some chance of being an extreme event. For extreme events, we simulated its stomach mass, $m_{ej}$ by $m_{ej} \sim \text{Gamma}(k_e, \theta_e)$, where $m_{ej}$ indicates the mass of the $j$th stomach is an extreme event.

We simulated data with these extreme events under multiple combinations of parameter values. We used the same variance for all stomachs (Table 2) with two values of $m_e | \phi$, $E[m_e | \phi] = 20$ and 50, and five probabilities that a stomach was an extreme event, $P(\text{extreme event}) = 0.01, 0.02, 0.05, 0.1, 0.2$.

Additionally, we simulated data that had more extreme outliers when $p=0$ (e.g. a very few particularly full stomachs when the predator had not eaten the prey type). This scenario was based on our analysis of predator stomach content data (below).

Specifically, here we simulated the case where stomachs that did not contain the prey type had a 0.2% chance of having a sample mass 50 times greater than the other stomachs, such that $E[m_e | \neg \phi] \sim 50*\text{Gamma}(k_a, 1/\theta_a)$. Again, we modified the basic operating model previously described. Data were simulated for $c_i = 0.3$ and 0.6.

Covariance Between Diet Fraction and Mass of Stomach Contents
We used simulated data to test the simplifying assumption that the mixture model makes about the relationship between total stomach contents’ mass of a predator and the proportion of the stomach contents that consists of a particular prey. Namely, the mixture model assumes that there are distinct stomach content masses when the prey item is present vs. absent, but that otherwise there is no relationship between diet fraction and stomach mass. Biologically, this implies that the total amount a predator consumes in a feeding episode depends solely on whether the predator has eaten the prey type in this feeding episode, and not how dominant the prey type is in the stomach. We sought to test the robustness of this assumption by generating data from an operating model in which there is a more complex relationship between $p_i$ and $m_s$. Specifically, we continued to use the basic operating model described above, except we assumed that when a stomach contained the prey type the mean stomach mass was a function of $p$:

$$E[M_{sj}] = E[M_{sj} | \Phi] + (\max\ consumption - E[M_{sj} | \Phi]) * P_j^x$$

We created linear covariance by setting $x = 1$, concave down covariance by $0 < x < 1$ and concave up covariance by $x > 1$ (Figure 2). For concave up and concave down covariance we tested a range of values for $x$.

**Applications**

We applied the mixture model, the conventional weighted mean, and conventional mean to stomach content data for four fish species with varying predation strategies and life histories. We obtained diet data for two small-bodied predators, Pacific herring, *Clupea pallasii*, and English sole, *Parophrys vetulus*, that have clear dominant prey types (euphausiids and polychaetes, respectively). We also used data for two large, generalist...
predators: Pacific spiny dogfish, *Squalus suckeyi*, and lingcod, *Ophiodon elongates*. All samples were collected in the Salish Sea, Washington, U.S.A. The stomach content data for Pacific herring, *Clupea pallasii*, were collected in Hood Canal, from June to October 2013 by midwater trawling (Sato et al, *in prep*). Both the English sole and dogfish were sampled throughout Puget Sound by bottom trawling over multiple seasons from 2004-2005 (Reum and Essington 2008). Finally, we also used diet information from lingcod collected individually in the San Juan Archipelago between spring 2004 and fall 2005 (Beaudreau and Essington 2009). As herring, English sole, and dogfish were caught in high numbers in single tows, only one stomach sample per trawl was used when estimating $E[c_i]$ to lower the likelihood of pseudoreplication. For each of the four datasets, we estimated the diet fractions of the most common prey and a prey type that exhibited covariance between $p_i$ and $m_s|\phi$.

RESULTS

Model Testing

We generated data from the basic operating model while varying the parameter values over their ranges to confirm our mixture model’s ability to accurately estimate $\hat{c}_i$ over the range of parameter values. The mixture model accurately estimated the true parameter values across the range of parameter values. The maximum mean absolute error (MAE) for all parameters was 0.018 (Table A1), indicating parameter estimates were robust over the ranges tested. The range of MAE for $\hat{c}_i$ was 0.004 to 0.015 and increased with the true value of $c_i$. 

https://mc06.manuscriptcentral.com/cjfas-pubs
The mixture model, the conventional mean, and conventional weighted mean differed in their sensitivity to overdispersion in the parameters controlling the frequency of the prey type’s occurrence, the frequency of $P_i=1$, and the distribution of individual diet fractions ($P_i$). The new mixture model and the conventional weighted mean accurately estimated $\hat{c_i}$ when there was overdispersion in any of these random variables, while the conventional mean underestimated $\hat{c_i}$, especially when the frequency of $P_i=1$ was overdispersed (Figure 3a-c). The precision (here, measured as the standard error of the estimate) of estimates was relatively similar across estimation methods and overdispersion scenarios, but tended to be slightly greater for the conventional sample mean (Figure 3d-f).

The estimation methods differed in their sensitivity to extreme events in total stomach contents’ mass. For all scenarios explored here (each varying the magnitude and frequency of samples with extreme stomach mass), the conventional mean underestimated diet fraction, while the mixture model and conventional weighted mean tended to accurately estimate diet fraction (Figure 4). The error in diet fractions estimated using a conventional mean was substantial, ranging from 0.08 – 0.12 (range of median values). The precision of estimates showed different patterns. Specifically, the conventional weighted mean generally had the lowest precision, especially when extreme events were uncommon, but of high magnitude (Figure 5). The precision of estimates...
from the mixture model were similar to, but slightly less than those of the conventional mean.

Covariance Between Diet Fraction and Mass of Stomach Contents

When the simulated data contained covariance between \( p_i \) and \( m_s \), the accuracy of estimates from the mixture model varied depending on the shape of covariance (Figure 6a). The mixture model was least accurate in the presence of linear covariance. When concave up or down covariance was present, the degree of concavity had little effect on the accuracy of the diet fraction estimate, such that the median underestimate of diet fraction was consistently about 0.05 (10% different from true value). In comparison, a conventional weighted mean was consistently accurate in estimating \( \bar{c}_i \), and a conventional mean was consistently more inaccurate than the mixture model. While these simulations tested positive covariance, we would expect the same patterns of bias, but in the opposite direction, for negative covariance.

The precision of the mixture model also varied depending on covariance shape and concavity (Figure 6b). With concave down covariance, the precision of the mixture model increased as the concavity of the relationship between \( p \) and \( m_s \) increased. The error from the mixture model was approximately equal to the error from a conventional weighted mean and conventional mean at \( x = 1/4 \) and was more precise at lower values for \( x \). However, for \( x \geq 1 \), the precision of the mixture model did not depend on the concavity and was consistently less precise than a conventional weighted mean or conventional mean.
Applications

For each dataset, we first estimated the diet fractions of the most common prey, as determined by the conventional mean. Estimates derived from the mixture model and the conventional weighted mean were broadly similar, but not identical (Figure 7). In three of the four applications, the estimates were as expected from the simulation analysis: the mixture model and conventional weighted mean estimates were similar, which were both very different from the conventional mean estimates, suggesting there was covariance between stomach total mass and diet fraction in the data. In the fourth application, Pacific herring, both the mixture model and conventional weighted mean estimates were much less than the conventional mean, though all estimates had large error (Figure 7a). In all applications, the precision of estimates derived from the mixture model and conventional weighted means was similar, and was generally less than that of the conventional mean (Figure 7).

In most instances when there was covariance between total stomach content mass and diet fraction, the mixture model produced estimates that were similar to those of the conventional weighted mean (Figure 8). This was counter to our expectations that followed from our simulation study. In particular, we expected that estimates from the mixture model would be intermediate to the highly biased conventional mean, yet this expectation was true in only one of the four cases (Figure 8b). The lack of similarities between the simulation study and applications to actual data likely reflects the facts that there was greater variance in the real data, and the patterns of covariance between $m_s$ and $p_i$ were not as strong as those simulated.
DISCUSSION

The mixture model represents a beneficial new method to quantitatively estimate prey contributions to predator diets using stomach content data. It advances the statistical modeling of these data by providing a robust, likelihood-based method that can be applied to complex data. The model was applied here to fish stomach content data, though this method is broadly applicable to other data types as well. The accuracy and precision of the mixture model was generally comparable to that of the conventional weighted mean approach. In most scenarios, the two methods performed similarly, and are both more accurate than using the simple average of individual diet fractions, as would be expected based on Ahlbeck et al. (2012). The mixture model performed particularly well compared to more conventional methods when there were a small number of very full stomachs in the sample set. Additionally, the mixture model does not require the user to make a decision on whether to weight stomachs by their stomach fullness or treat them as equally informative, which has traditionally been a challenge when working with stomach content data (Nielsen et al. 1983). While the model will produce biased estimates when there is a strong pattern of covariance between sample stomach mass and diet fraction, applications to real data did not indicate that this scenario was common.

The introduction of a likelihood framework into diet estimation methods greatly increases the range of uses for diet data. Analysis methods that include likelihood frameworks have been developed for count data (de Valpine and Harmon-Threatt 2013; Fletcher et al. 2005; Lynch et al. 2014), but this mixture model is explicitly created for proportional data. Formal model selection is standard in other areas such as regression (Burnham and Anderson 1998; Hilborn and Mangel 1997) and is typically done using
AIC or likelihood ratio tests. Both of these can now be carried out for diet data with the mixture model framework. Additionally, the existence of a likelihood function provides the ability to use Bayesian methods (Gelman and Hill 2007), which permits various model extensions (e.g., including prior information based on stable isotope analysis or previous stomach content analysis). For instance, one possibility is extending the mixture model to develop a hierarchical model to account for lack of independence among samples. Individual stomachs captured in a single sampling event are frequently not independent samples (Nielsen et al. 1983). A hierarchical-based extension of our likelihood framework would allow one to directly estimate the level of independence rather than making extreme assumptions (i.e. all samples in a collection event are wholly independent or dependent) (Gelman and Hill 2007).

When data includes stomachs with large stomach masses, analysts would typically choose between a conventional weighted mean or conventional mean by balancing benefits and costs of the two methods. As an example, consider the case of a single stomach that contains one large prey item whose mass represents a substantial fraction of the total sampled mass. Previously, in such a case, analysts have had to choose whether to pool samples (or equivalently use a conventional weighted mean), or to use a conventional sample mean (Nielsen et al. 1983). The former would clearly overstate the importance of that prey item, while the latter is known to impose a substantial bias (Ahlbeck et al. 2012). With the mixture model, the analyst does not need to make this choice, because the estimation procedure is highly robust to these events that often complicate stomach content analysis.
Simulation testing indicated potential bias of the mixture model when the relationship between stomach fullness and stomach contents was more nuanced than the model assumed. We also analyzed stomach contents for the four predator types in cases where the conditions for bias were most likely to occur. In three of these four cases, there was little evidence of bias in the mixture model. We did not observe the mixture model estimating a diet fraction intermediate to conventional mean and conventional weighted mean estimates. This implies that in many cases, the conditions that gave rise to bias in simulations may be uncommon in real datasets. A careful examination of diet data for covariance between \( p_i \) and \( m_s \), and comparison of estimates from the mixture model to other methods, can reveal the potential for bias in the mixture model. An initial examination of the diet data should, however, be part of any analysis (Burnham and Anderson 1998; Gelman and Hill 2007; Hilborn and Mangel 1997).

The model framework can also be adapted to minimize the potential bias caused by the model simplification of estimating two discrete stomach masses. The choice to have two distinct stomach content parameters for stomachs that do and do not contain prey type \( i \), \( m_s|\Phi \), when \( p_i > 0 \), and another stomach mass when the prey is absent, \( m_s|\neg\Phi \), when \( p_i = 0 \), allows us to partially address the covariance between \( p_i \) and \( m_s \). However, there is no constraint on the number of distinct prey mass distributions that can be assumed, and including more prey masses could remove much of this bias. For example, one could estimate a distinct conventional mean stomach mass for stomachs when \( p_i = 1 \), or divide the mass data into multiple cases based on the observed relationship between \( p_i \) and \( m_s \). A continuous functional relationship between \( E[m_s] \) and \( p_i \) could also be assumed to avoid making this assumption at all.
Most methods for estimating diet fractions only estimate the contribution of a single prey type to the predator’s diet, rather than estimating the predator’s entire diet composition simultaneously (Ahlbeck et al. 2012; Liao et al. 2001). If this is done for each prey type, the resulting estimates will not sum to precisely 1. A multivariate beta distribution, known as the Dirichlet distribution, could be used to estimate the diet fractions for all consumed prey types at once. In fact, this was the method employed by Ainsworth et al. (2010) to combine estimates from multiple data sources. However, the Dirichlet cannot be used to estimate diet contributions if any of the samples contain $p_i=0$ or $p_i=1$, i.e. each prey type needs to be present in each sample. Standard application of the Dirichlet also weights all samples identically regardless of stomach fullness.

The new model is an important step forward in quantitatively estimating predator diets and is also broadly applicable to other data types involving proportional data. There are, of course, limitations to diet data that are not addressed by our method. These include differential digestion (Baker et al. 2014), regurgitation during the sampling process, and small sample sizes that may not capture rare, but ecologically important, events. However, the mixture model opens the door to addressing other challenges of proportional stomach content data by providing a likelihood framework for diet fractions estimates, while generally providing accurate and precise estimates of diet fractions.

ACKNOWLEDGMENTS

P. Moriarty was supported by a Department of Defense National Defense Science and Engineering Graduate (NDSEG) Fellowship, T. Essington was supported by National
Science Foundation grant OCE-1154648AM001. We thank Lewis Barnett, Emma Hodgson, John Horne, Julie Keister, Laura Koehn, Kristin Marshall, Kiva Oken, Margaret Siple and two anonymous reviewers for manuscript comments.
References


Marshall, K.N. 2007. Integrating energetics and interaction strengths in natural ecosystems. MSc Thesis., School of Aquatic and Fishery Sciences, University of Washington, Seattle, WA.


### Table 1
The parameters in the mixture model, along with their corresponding distributions and meanings.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Distribution</th>
<th>Parameter Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$c_i$</td>
<td>-</td>
<td>Mean contribution of prey type $i$ to a predator’s diet</td>
</tr>
<tr>
<td>$P_i$</td>
<td>$\beta(\alpha_1,\alpha_2)$</td>
<td>Mean fraction of the predator’s diet made up of prey type $i$</td>
</tr>
<tr>
<td>$M_s</td>
<td>\Phi$</td>
<td>$\Gamma(k_{\text{present}}, 1/\theta_{\text{present}})$</td>
</tr>
<tr>
<td>$M_s</td>
<td>\neg\Phi$</td>
<td>$\Gamma(k_{\text{absent}},1/\theta_{\text{absent}})$</td>
</tr>
<tr>
<td>$\Phi$</td>
<td>Ber($r_\Phi$)</td>
<td>Frequency of occurrence of prey type $i$ in samples</td>
</tr>
<tr>
<td>$\Phi_{p_1=1}$</td>
<td>Ber($r_{\Phi p_1=1}$)</td>
<td>Frequency a sample only contains prey type $i$ given it contains prey type $i$</td>
</tr>
<tr>
<td>$\sigma_\Phi$</td>
<td>-</td>
<td>Standard deviation of $\Phi$</td>
</tr>
<tr>
<td>$\sigma_{\neg\Phi}$</td>
<td>-</td>
<td>Standard deviation of $\Phi_{p_1=1}$</td>
</tr>
<tr>
<td>$\sigma_i$</td>
<td>-</td>
<td>Standard deviation of $P_i$</td>
</tr>
</tbody>
</table>

### Table 2
True values used in sensitivity scenarios for parameters that were held constant.

| Scenario                  | $r_\theta$ | $r_{0p=1}$ | $m_s|\theta$ | $\sigma_{ms|\theta}$ | $\sigma_{ms|\neg\theta}$ | $p_i$ | $\sigma_p$ | $c_i$ |
|---------------------------|------------|------------|--------------|----------------------|--------------------------|-------|------------|-------|
| Overdispersion            | -          | -          | 4            | 1                    | 1                        | -     | 0.1        | 0.5   |
| Extreme Events            | 0.5        | 0.5        | 4            | 1                    | 1                        | 0.5   | 0.1        | -     |
| Covariance Between $p_i$ and $m_s$ | 0.5 | 0.5 | - | 0.1 | 0.1 | 0.5 | 0.4 | 0.5 |
Figure 1. Conceptual flowchart of the mixture model, which splits stomachs into the three cases of $p_i = 0$, $0 < p_i < 1$ and $p_i = 1$. The parameters relevant to each case are shown along with the distributions used to estimate them in the mixture model.

Figure 2. The continuous relationship between $E(M_s)$ and $P_i$, which can take a range of shapes, is simplified into a discrete relationship of presence / absence in the mixture model. During model testing, data with concave up, concave down and linear covariance between $p_i$ and $m_s|\Phi$ was generated with varying degrees of concavity to test the mixture model’s robustness to this simplification. The degree of concavity was altered by varying the value of $x$ in the power function, $E[M_s] = E[M_s|\Phi] + (\text{max consumption} - E[M_s|\neg\Phi]) \times P_f^x$.

Figure 3. Effect of overdispersion in probability of prey occurrence ($R_{\Phi}$; a, d), probability of only that prey occurring given it does occur ($R_{\Phi p=1}$; b, e), and diet fraction when the prey is present ($P_{i.}$; c, f) on the accuracy (top) and standard error (bottom) of diet fraction estimates. Vertical line shows the true diet fraction. Results are based on 100 simulated data sets each consisting of 200 samples. Box plots indicate median, interquartile range (boxes) and 150% interquartile range (lines).

Figure 4. Effect of extreme events in stomach masses (e.g., a few very full stomachs) on the accuracy of diet fraction estimates across multiple values of stomach mass when the prey type is present, $((a-c) \ M_s|\Phi = 20, (d-f) \ M_s|\Phi = 50)$ and across multiple values of the probability of an extreme event occurring, $((a,d) \ P(\text{extreme event}) \ = 0.01, (b,c) \ P(\text{extreme event}) \ = 0.05, (c,f) \ P(\text{extreme event}) \ = 0.1)$. Scenarios for $P(\text{extreme event}) = 0.02, 0.2$ were also explored and results were consistent with the plots shown here. Horizontal lines show the true value of $c_i$. Each box represents 100 simulated datasets of 200 stomachs. Box plots indicate median, interquartile range (boxes) and 150% interquartile range (lines).

Figure 5. Effect of extreme events in consumption (e.g., a few very full stomachs) on the precision of diet fraction estimates across multiple values of stomach mass when the prey type occurs $((a-c) \ M_s|\Phi = 20, (d-f) \ M_s|\Phi = 50)$ and across multiple values of the probability of an extreme event occurring $((a,d) \ P(\text{extreme event}) \ = 0.01, (b,c) \ P(\text{extreme event}) \ = 0.05, (c,f) \ P(\text{extreme event}) \ = 0.1)$. Scenarios for $P(\text{extreme event}) = 0.02, 0.2$ were also explored and results were consistent with the plots shown here. Horizontal lines show the true value of $c_i$. Each box represents 100 simulated datasets of 200 stomachs. Box plots indicate median, interquartile range (boxes) and 150% interquartile range (lines).

Figure 6. Effect of covariance between diet fraction $P_i$ and total stomach contents’ mass on the accuracy and precision of the mixture model, a convention mean (dashed line) and conventional weighted mean (dotted line). Varying degrees of concavity in the data were created by using a power function, $E[M_s] = E[M_s|\neg\theta] + (\text{max consumption} - E[M_s|\neg\Phi]) \times P_f^x$. 
$E[M_{sj} | \theta] \ast P_j^x$ and varying $x$. Each box represents 100 simulated datasets of 200 stomachs. Box plots indicate median, interquartile range (boxes) and 150% interquartile range (lines).

Figure 7. Estimates and standard error when the mixture model, a conventional mean, and conventional weighted mean are applied to a common prey type in diet datasets for four predators with varying life histories, a) Pacific herring, *Clupea pallasii*, b) English sole, *Parophrys vetulus*, c) lingcod, *Ophiodon elongates*, and d) spiny dogfish, *Squalus acantbias*. Error bars are standard error.

Figure 8. Estimates and standard error when the mixture model, a conventional mean, and conventional weighted mean are applied to a prey type demonstrating correlation between stomach mass and diet fraction. All three methods are applied to diet datasets for four predators with varying life histories, a) herring consuming glass shrimp, b) English sole consuming amphipods, c) spiny dogfish consuming Pacific herring, and d) lingcod consuming great sculpin. Error bars are standard error. Insets are actual data with regression lines, plotted as total stomach contents’ mass versus diet fraction.
Figure 1. Conceptual flowchart of the mixture model, which splits stomachs into the three cases of $p_i = 0$, $0 < p_i < 1$ and $p_i = 1$. The parameters relevant to each case are shown along with the distributions used to estimate them in the mixture model.
Figure 2. The continuous relationship between $E(M_s)$ and $P_i$, which can take a range of shapes, is simplified into a discrete relationship of presence/absence in the mixture model. During model testing, data with concave up, concave down and linear covariance between $p_i$ and $m_s|\phi$ was generated with varying degrees of concavity to test the mixture model’s robustness to this simplification. The degree of concavity was altered by varying the value of $x$ in the power function, $E[M_{sj}] = E[M_{sj}|\phi] + (\text{max consumption} - E[M_{sj}|\neg\phi]) * P_i^x$. 
Figure 3. Effect of overdispersion in probability of prey occurrence ($R_{φ}$; a, d), probability of only that prey occurring given it does occur ($R_{φp} = 1$; b, e), and diet fraction when the prey is present ($P_i$; c, f) on the accuracy (top) and standard error (bottom) of diet fraction estimates. Vertical line shows the true diet fraction. Results are based on 100 simulated data sets each consisting of 200 samples. Box plots indicate median, interquartile range (boxes) and 150% interquartile range (lines).
Figure 4. Effect of extreme events in stomach masses (e.g., a few very full stomachs) on the accuracy of diet fraction estimates across multiple values of stomach mass when the prey type is present, ((a-c) $M_\| \Phi = 20$, (d-f) $M_\| \Phi = 50$) and across multiple values of the probability of an extreme event occurring, ((a,d) $P(\text{extreme event}) = 0.01$, (b,e) $P(\text{extreme event}) = 0.05$, (c,f) $P(\text{extreme event}) = 0.1$).

Scenarios for $P(\text{extreme event}) = 0.02$, 0.2 were also explored and results were consistent with the plots shown here. Horizontal lines show the true value of $c_i$. Each box represents 100 simulated datasets of 200 stomachs. Box plots indicate median, interquartile range (boxes) and 150% interquartile range (lines).
Figure 5. Effect of extreme events in consumption (e.g., a few very full stomachs) on the precision of diet fraction estimates across multiple values of stomach mass when the prey type occurs ((a-c) $M_s|\Phi = 20$, (d-f) $M_s|\Phi = 50$) and across multiple values of the probability of an extreme event occurring ((a,d) $P(\text{extreme event}) = 0.01$, (b,e) $P(\text{extreme event}) = 0.05$, (c,f) $P(\text{extreme event}) = 0.1$). Scenarios for $P(\text{extreme event}) = 0.02, 0.2$ were also explored and results were consistent with the plots shown here. Horizontal lines show the true value of $c_i$. Each box represents 100 simulated datasets of 200 stomachs. Box plots indicate median, interquartile range.
Figure 6. Effect of covariance between diet fraction $P_i$ and total stomach contents’ mass on the accuracy and precision of the mixture model, a convention mean (dashed line) and conventional weighted mean (dotted line). Varying degrees of concavity in the data were created by using a power function, $E[M_{sj}] = E[M_{sj}|-\theta] \pm (\text{max consumption} - E[M_{sj}|-\theta]) \times P_j^x$ and varying $x$. Each box represents 100 simulated datasets of 200 stomachs. Box plots indicate median, interquartile range (boxes) and 150% interquartile range (lines).
Figure 7. Estimates and standard error when the mixture model, a conventional mean, and conventional weighted mean are applied to a common prey type in diet datasets for four predators with varying life histories, a) Pacific herring, *Clupea pallasii*, b) English sole, *Parophrys vetulus*, c) lingcod, *Ophiodon elongates*, and d) spiny dogfish, *Squalus acanthias*. Error bars are standard error.
Figure 8. Estimates and standard error when the mixture model, a conventional mean, and conventional weighted mean are applied to a prey type demonstrating correlation between stomach mass and diet fraction. All three methods are applied to diet datasets for four predators with varying life histories, a) herring consuming glass shrimp, b) English sole consuming amphipods, c) spiny dogfish consuming Pacific herring, and d) lingcod consuming great sculpin. Error bars are standard error. Insets are actual data with regression lines, plotted as total stomach contents’ mass versus diet fraction.
### Appendix A

Table A1. Mean absolute error (MAE) when each parameter in the mixture model was varied over a range of values. The minimum and maximum MAE do not necessarily correspond to the minimum and maximum parameter values.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Range</th>
<th>MAE Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>r_\phi</td>
<td>0.1-0.9</td>
<td>0.021-0.089</td>
</tr>
<tr>
<td>r_{op=1}</td>
<td>0.1-0.9</td>
<td>0.024-0.034</td>
</tr>
<tr>
<td>m_{\phi}</td>
<td>1.0-10</td>
<td>0.027-0.036</td>
</tr>
<tr>
<td>c_i</td>
<td>0.1-0.6</td>
<td>0.012-0.035</td>
</tr>
<tr>
<td>\sigma_\phi</td>
<td>0.1-2</td>
<td>0.027-0.032</td>
</tr>
<tr>
<td>\sigma_-\phi</td>
<td>0.1-2</td>
<td>0.026-0.031</td>
</tr>
<tr>
<td>p</td>
<td>0.1-0.9</td>
<td>0.027-0.036</td>
</tr>
<tr>
<td>\sigma_p</td>
<td>0.01-0.4</td>
<td>0.026-0.030</td>
</tr>
</tbody>
</table>