Replacing the multinomial in stock assessment models: A first step

R.I.C. Chris Francis

2014

SEDAR56-RD13

9 January 2018

Volume 151, March 2014
ISSN 0165-7836
ELSEVIER
Fisheries Research
An international journal on fisheries science, fishing technology and fisheries management

Fisheries Research 151 (2014) 70–84

Contents lists available at ScienceDirect

Fisheries Research

journal homepage: www.elsevier.com/locate/fishres

Replacing the multinomial in stock assessment models: A first step

R.I.C. Chris Francis*

123 Overtoun Terrace, Wellington 6021, New Zealand

ARTICLE INFO

Article history:
Received 23 October 2013
Received in revised form
12 December 2013
Accepted 15 December 2013

Keywords:
Fisheries stock assessment
Composition data
Multinomial likelihood
Logistic-normal likelihood
Dirichlet likelihood

ABSTRACT

Though it is by far the most commonly used likelihood for composition data (proportions at length or age) in fisheries stock assessment models, the multinomial is poorly suited for this task. It has two salient weaknesses: it can not replicate the correlations found in these data; and it is not self-weighting (i.e., the parameters that weight the composition data can not be estimated inside the model). This latter weakness derives from the fact that the multinomial likelihood, being designed for discrete data but used for continuous data, is improper (i.e., its integral over all permissible data values is not constant). All other likelihoods commonly used for composition data share at least one of these weaknesses but there is one – the logistic-normal – which can be extended to avoid both. Some, like the multivariate normal, are misused because their structure ignores the defining properties of composition data: that they lie between 0 and 1, and sum to 1. A collection of 72 composition data sets from 28 stock assessments originating from nine different computer programs was used to evaluate the extended logistic-normal, together with the Dirichlet likelihood, which is self-weighting but does not allow positive correlations (and so may be useful for composition data with small correlations). The logistic-normal appears very promising, especially for unsexed length compositions. The next step in evaluating the extended logistic-normal likelihood will be to code it into stock assessment programs, and some of the technical problems associated with this step are discussed.

1. Introduction

The provision of scientific advice for the management of major fisheries involves the analysis of all available data in what are called stock assessment models. These models estimate the likely exploitation history of the fish stock (how many fish there were of each age in each year, and what proportion of these were caught) in order to answer management questions about productivity and sustainability. In addition to annual catches, the main two types of data analysed by these models are abundance indices (from surveys or fishery catch rates) and *compositions*, which describe the distribution of lengths or ages in catches (from surveys or fisheries) in each year. The latter type is important in estimating mortality rates and year-to-year variations in the strength of recruitment to the population. In modern assessment models, estimation is likelihood-based (Maunder and Punt, 2013). That is to say, for each data set analysed by the model there is a mathematical expression, called a *likelihood*, which measures how consistent that data set is with any potential exploitation history. It does that by providing a statistical description of the assumed error structure of the data set (the likely difference between the data values and those expected by the model). The choice of likelihoods in a model is important because changing likelihoods can substantially change key model outputs. This happens in part because a change of likelihoods will change the relative weights assigned to individual data sets, and also to individual data points within a data set (see text associated with Fig. 1 of Francis, 2011, for an example of how data weighting can change model outputs). Moreover, valid statistical inferences from these models (e.g., confidence intervals for estimates, or use of AIC (Akaike, 1974) to decide between competing versions of the model) require the use of appropriate likelihoods (Deriso et al., 2007).

This paper addresses the question of what is the best likelihood to use for composition data in stock assessment models. The history of this usage is relatively brief. Before computers became sufficiently powerful to fit models with many parameters, early assessment methods (such as Virtual Population Analysis (e.g., Pope, 1972; Shepherd, 1999)) needed no likelihoods for compositions because these data were assumed to be without error. In the early years of statistical stock assessment models these data were fitted by least squares in log space (e.g., Deriso et al., 1985; Kimura, 1989). This implied that the error structure was such that all proportions at age (or length) had the same coefficient of variation (c.v.). Crone and Sampson (1998) showed that this was not true: c.v.s declined with increasing proportions, as is characteristic of the multinomial distribution (for which the c.v. of a proportion

* Tel.: +64 4 386 3517.
E-mail address: chris.francis@clear.net.nz

0165-7836/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.fishres.2013.12.015

with expected value p and sample size N is $[(1-p)/(pN)]^{0.5}$). In recent years, the multinomial has become by far the most common likelihood used for composition data. For example, Stock Synthesis (Methot and Wetzel, 2013), which is arguably the most widely used general-purpose statistical stock assessment program, provides no other likelihood for composition data.

There are two well known problems with the use of the multinomial likelihood for fishery composition data: *overdispersion* and *correlation* (Hrafnkelsson and Stefánsson, 2004). Both arise from the fact that the raw data from which a single composition is calculated do not, as is assumed by the multinomial distribution, consist of a single simple random sample from the catch which is to be characterised by that composition (usually the annual catch from a fishery, or the total catch from a survey). Rather, they are random samples from many individual tows or sets. It is widely recognised that these data show what Pennington and Vølstad (1994) called *intra-haul correlation* – fish caught in the same tow or set are more like each other (in length or age) than fish from different tows or sets. As a consequence, to obtain appropriate c.v.s for composition data, the sample size parameter, N, in the multinomial likelihood must be set much smaller than the actual sample size (this addresses the overdispersion problem). Many different ways have been used to make this correction for overdispersion (e.g., Crone and Sampson (1998) constructed an empirical relationship between the corrected sample size and the number of trips sampled [see their Fig. 4]; McAllister and Ianelli (1997) devised an algorithm to correct initial sample sizes using the output from a run of the stock assessment model [see eqs. (2.5) and (2.6) in their appendix 2]). The problem of correlation is harder to deal with. The intra-haul correlation induces correlations between the individual proportions in a composition (e.g., the proportion of fish of length 20 cm in a composition is correlated with the proportion of length 21 cm). These correlations are often substantial, particularly for length compositions (e.g., Hrafnkelsson and Stefánsson, 2004, Fig. 4; Miller and Skalski, 2006, Figs. 2–6). Pennington and Vølstad (1994) provided a striking demonstration of the effect of these correlations. They calculated, from fish caught in a trawl survey, the standard error (s.e.) of the estimated mean length of fish in the survey area, and defined the *effective sample size* for the length composition from that survey as being the sample size that would be required to produce that s.e. if we were able to take a simple random sample from the total survey catch. These effective sample sizes (which I shall call *Pennington sample sizes*, denoted N_{Penn}) can be surprisingly small. For example, for haddock length compositions from a series of trawl surveys on Georges Bank, Pennington and Vølstad calculated a median effective sample size of 21 (range 3–152), which was about half the number of tows that caught haddock (median 41, range 22–124) and much less than the number of fish measured (median 845, range 157–12 208). In the stock assessment setting, the problem is that these substantial correlations, both positive and negative, are inconsistent with the multinomial likelihood, for which correlations are usually small and always negative (the correlation between multinomial proportions with expected values p_b, p_c in bins b and c is $-[p_b p_c/\{(1-p_b)(1-p_c)\}]^{0.5}$). Thus it is not possible with the multinomial likelihood to include realistic correlations for composition data.

Francis (2011) pointed out that the effect of correlations is to reduce the amount of information in composition data, and thus the weight that should be given to them. He suggested that there were three ways to deal with these correlations: (1) ignore them (the common response); (2) discard the multinomial in favour of a likelihood which allows substantial correlations; or (3) reduce the multinomial sample size to compensate for the correlations. He said that option 1 was unsatisfactory because it tended to over-weight the composition data sets, which can cause poor fits to abundance data. Since he was unaware of a likelihood suitable for option 2, he recommended option 3, and proposed a method to implement it using Pennington sample sizes. In this paper I pursue option 2. I will list the properties ideally found in a composition likelihood and show that though none of the currently used likelihoods has all of these properties there is one (the logistic-normal) which can be extended to do so. I will then analyse composition data from a wide range of stock assessments to see to what extent they support this extended likelihood (R functions used in these analyses are provided as Supplementary Data). Before starting on this path I need to discuss different types of error.

2. Types of error

It might seem that the obvious place to look, when seeking a likelihood which will faithfully represent the error structure of composition data, is in the results from recent studies which have analysed the raw data contributing to a composition (e.g., Kvist et al., 2001; Rindorf and Lewy, 2001; Hirst et al., 2004; and others mentioned above). This is only partly true. We can't use these results directly because they concern a type of error which differs from that which we are addressing. To understand the difference, consider an individual composition proportion that may of interest in a stock assessment (say the proportion at age 2 in the catch from the longline fishery in 2012). As Francis (2011) has noted, there are three different versions of that proportion: (i) the value we observe, O; (ii) the true (real world) value, T; and (iii) the value expected by our stock assessment model, E. The studies just mentioned tell us about the *observation error*, which is the difference between O and T. However, the error we wish to represent in our composition likelihood is $(O-E)$. Francis called this the *total error* because it is the sum of the observation error, $(O-T)$, and $(T-E)$, which he called the *process error* (this arises because the many simplifying assumptions that are needed to make the model tractable mean that the stock assessment model is only an approximation to the truth). Thus the variance, or c.v., of the total error must be greater than that of the observation error. We could treat the observation error as being a reasonable approximation to the total error only if we knew that the process error was small compared to the observation error. My experience is that this is unlikely to be true for composition data, and I offer an example to illustrate the point.

The data for this example come from the 2012 assessment of hoki (*Macruronus novaezelandiae*) in New Zealand (McKenzie, 2013). This assessment modelled two stocks of hoki which have separate spawning grounds but a common nursery ground. The main composition data sets used were proportions at age from the fisheries on the spawning grounds, HOKwc and HOKcs, and a survey on the nursery ground, HOKcr. By bootstrapping the raw data (length samples, and ages for age-length keys) we can calculate the s.e. of mean age for each composition, and thus the associated Pennington sample size. These sample sizes, which characterise the size of the observation error for each data set, varied from year to year, but their median values were of a similar order to (for HOKwc) or less than (for HOKcs and HOKcr) the number of otoliths sampled, and much less than the number of lengths (Table 1). We can also calculate N_{Penn} values for the total error from the model residuals (i.e., $O-E$) as described by Francis (2011) (see method TA1.8 in his appendix A). This was done both for the base model (labelled 1.3 in the assessment), in which the composition data were sexed (i.e., they consisted of proportions by age and sex), and also for an alternative model (labelled 1.4) which used exactly the same composition data but without sex. The total-error sample sizes were markedly smaller than those for observation error, particularly for the two fishery data sets (Table 1). Thus the process error is not negligible for these data sets. One obvious source

Table 1
Various sample sizes calculated for three age composition data sets used in the 2012 assessment of hoki (*M. novaezelandiae*). Columns 3–5 are medians of values calculated by year for each data set, and contain the numbers of otoliths and fish lengths sampled, N_{otolith} and N_{length}, and the Pennington sample sizes, N_{Penn}, for observation error. Columns 6–7 are the Pennington sample sizes for total error, calculated from two alternative models: one in which the composition data were sexed, and the other in which they were not.

		Median numbers from observations			N_{Penn} for total error	
Data set	Type	N_{otolith}	N_{length}	N_{Penn}	Sexed	Unsexed
HOKwc	Fishery	753	53 061	937	20	27
HOKcs	Fishery	763	10 527	261	69	69
HOKcr	Survey	649	19 172	116	83	52

of process error is the assumption that selectivities did not vary between years. The fact that process error was smaller for HOKcr could be because this assumption is likely to be closer to the truth for a survey than for fisheries. Assumptions about natural mortality are another source of process error for composition data. Natural mortality was assumed to be time-invariant in both models; in the alternative model it was also assumed to be independent of age.

3. Composition likelihoods: notation, desirable properties and examples

For a given composition data set, let O_{by} denote the observed proportion in bin b (for $b = 1, \ldots, B$) and year y (for $y = 1, \ldots, Y$), where the bins are either age or length classes, and let E_{by} denote the corresponding expected value from the assessment model. Since these are compositions, they must sum to 1 in each year (i.e., $\Sigma_b O_{by} = \Sigma_b E_{by} = 1$ for all y). I will sometimes use vector notation, denoting the composition for year y as $\mathbf{O}_y$, and its expected value as $\mathbf{E}_y$; when referring to a single composition, I may, for simplicity, drop the subscript y.

A composition likelihood is a mathematical formula which may be thought of as measuring how consistent the model's expected values, E_{by}, are with the observations, O_{by}. The bigger the calculated likelihood is, the more consistent the expected values are with the observations (i.e., the better the model fits the data). By convention, assessment models calculate the negative logarithm of the likelihood, NLL, rather than the likelihood, so the smaller the NLL, the better the fit to the data. Technically, the E_{by} are parameters of the likelihood, which are always estimated in the model. Each likelihood will also have other parameters, which I will call *weighting parameters*, and which may be fixed (and adjusted) outside the model, or estimated inside it. For example, with the multinomial likelihood the NLL is given by $\sum_y \log\left(N_y!\right) + \sum_{by} \log\left[\left(N_y O_{by}\right)!\right] - \sum_{by}\left(N_y O_{by} \log E_{by}\right)$, where the weighting parameters, the N_y, are called the multinomial sample sizes for each year.

What properties are desirable for a composition likelihood? I shall list four. One obvious one, given the above discussion, is that it must allow substantial correlations. Another is that it should be *self-weighting*, by which I mean that we should be able to estimate the weighting parameters within the model. Not all likelihoods are self-weighting, and for those that are not we have two ways of dealing with the weighting parameters: fix them outside the model; or iteratively adjust them. The first approach is not advisable because the values we use should be appropriate for the total error in the composition data, and we have no way of knowing about the extent of that error without running the model and examining the residuals. The second approach is that commonly used with the multinomial likelihood in Stock Synthesis (Methot and Wetzel, 2013), which outputs information that can be used to iteratively adjust the values of N_y. This approach is workable, though inconvenient (because, at least theoretically, we need to adjust the sample sizes every time we make some change to the model assumptions or inputs). Also, there is some doubt about which is the correct algorithm to use to adjust the N_y (see the results of four alternative algorithms in the example in Table 4 of Francis, 2011). Self-weighting likelihoods are much more convenient, and avoid uncertainty about weighting algorithms. Another desirable property is that a composition likelihood should be *proper*. That is, for any given values of its parameters, its integral over all permissible values of the observations must be independent of the parameter values (this definition is a slightly less strict than the usual one, which requires that the integral equal 1). This property is relevant, though not essential, simply because a likelihood can not be correctly self-weighting unless it is proper; an improper likelihood will cause bias in estimates of weighting parameters. In principle, we can make any likelihood proper by adding (to the NLL) a term derived from the integral of the likelihood. However, this will be practicable only if that integral has a closed form. Finally, it is desirable that our likelihoods be *parsimonious* (i.e., have few weighting parameters). This is worth mentioning because some researchers have constructed models with large numbers of parameters for the observation error in composition data (e.g., Hrafnkelsson and Stefánsson, 2004; Miller and Skalski, 2006). These models are quite appropriate in the settings in which they were constructed because there was a lot of data (i.e., all the raw data that are used to construct the composition) from which to estimate the many parameters. They are not appropriate for a stock assessment model, which is fitted to the compositions, rather than the raw data from which they are constructed, and thus does not allow the estimation of many weighting parameters.

3.1. Which likelihoods have these desirable properties?

So, how do likelihoods that have been used for composition data perform with regard to these desirable properties? I will discuss seven likelihoods (Table 2). First, the multinomial. As noted above, it does not allow for substantial correlations, but it is certainly parsimonious (with just one weighting parameter per composition, N_y). The fact that it is not proper may surprise some readers. Since the multinomial distribution appears in all standard statistics texts one might expect its likelihood to be proper. The point to notice is that it appears in these texts as a discrete distribution, defined only for observed proportions that are multiples of $1/N$, and that its likelihood is proper when restricted to these values. In a stock assessment setting we use it as a continuous distribution, allowing the observed proportions to take any value between 0 and 1. The integral of the multinomial likelihood over this continuous domain does not have a closed form, but it is easy to show, by approximation, that it decreases as N increases, which is why attempts to estimate N within the model fail (the estimates always tend to zero). The multivariate normal likelihood was adapted for use with composition data – first in the modal analysis program MULTIFAN (Fournier et al., 1990), and subsequently in the stock assessment program MULTIFAN-CL (Fournier et al., 1998) – by reparameterisation (replacing the usual variance parameters with sample sizes, N_y, in such a way that the variances are the same as for the multinomial – see Eq. (A17) in the Appendix) and robustification. This likelihood has proved very useful in tuna stock assessments, which

Table 2
Some likelihoods that have been used for composition data, and their performance with respect to four desirable properties. The entries 'Possibly' and 'Usually' mean that the performance depends on how the likelihoods are implemented: 'Possibly' indicates that substantial correlations are possible, but not in current implementations; 'Usually' means that the likelihood is parsimonious in current implementations, but may not always be. 'Partially' indicates that the likelihood is not proper or self-weighting relative to all weighting parameters.

Likelihood	Substantial correlations?	Self- weighting?	Proper?	Parsimonious?
Multinomial	No	No	No	Yes
Multivariate normal	Possibly	No	No	Usually
Multivariate lognormal	Possibly	No	No	Usually
Logistic-normal	Possibly	Yes	Yes	Usually
Dirichlet	No	Yes	Yes	Yes
Dirichlet-multinomial	No	Partially	Partially	Yes
Logistic-multinomial	Possibly	Partially	Partially	Usually

are characterised by large quantities of noisy length composition data. However, its use illustrates a point made by Aitchison (2003). Writing about a broad range of types of statistical inference from general composition data (not just fishery proportions at age or length in stock assessment models), he noted that many published analyses of composition data lack rigour because they use general-purpose statistical tools (e.g., ANOVA) which ignore the special properties of these data (i.e., that they are non-negative proportions that must sum to one). This criticism applies to the multivariate normal likelihood (as used in stock assessment models), which does not constrain observations to be between 0 and 1, and to sum to 1. When these constraints are applied, the integral of the likelihood depends on the N_y, and so is neither proper nor correctly self-weighting. As used in MULTIFAN-CL, it is parsimonious (having the same parameters as the multinomial), but allows no correlations at all. It would be easy to modify it to allow substantial correlations, but care would need to be taken to do this in a parsimonious way (the most general multivariate normal likelihood would have $0.5YB(B-1)$ weighting parameters). The multivariate lognormal likelihood (e.g., Punt and Kennedy, 1997; Bull et al., 2012) has, with regard to our four desirable properties, exactly the same strengths and weaknesses as the multivariate normal. These strengths and weaknesses are unchanged in alternative versions of the normal and lognormal likelihoods, preferred by some researchers, in which variances depend on O_{by} rather than E_{by} (see, e.g., methods Fobs and PKobs in Table 1 of Maunder, 2011).

The first use of a stock assessment composition likelihood that was designed for continuous composition data (and was thus both proper and correctly self-weighting) was of the *logistic-normal* likelihood in a theoretical state-space model by Schnute and Richards (1995). (They called their likelihood "multivariate logistic", but Aitchison (2003), who developed much of the theory surrounding this distribution, calls it the (additive) logistic-normal). This likelihood is implemented in the (non-state-space) stock assessment program iSCAM (Martell, 2011; Martell et al., 2011). A logistic-normal distribution is formed by applying a logistic transformation to a multivariate normal vector. Specifically, a composition **O** is logistic-normal with parameters {**E**, **C**} if $O_b = \exp(X_b)/\sum_{b'}\exp(X_{b'})$, where **X** is multivariate normal with mean log(**E**) and covariance matrix **C** (here I am following the approach of Schnute and Haigh (2007), which differs slightly from that of Aitchison (2003), but is better suited to the stock assessment setting – see Appendix for details). Note that the logistic transformation forces the O_b to be positive proportions summing to 1. In the just-cited stock assessment applications of this distribution the covariance matrix **C** took its simplest possible form with $V(X_b) = \sigma^2$ for all b, and all covariances set to 0. This version of the logistic-normal is certainly parsimonious (it has only one weighting parameter, σ) but does not allow for substantial positive correlations. A simple and parsimonious way to introduce correlations is to set $\mathrm{Cor}(X_b, X_c) = AC_k(|b-c|)$, where b and c are bin numbers and AC_k is the auto-correlation function of a k-th order autoregressive process, AR(k) (Brockwell and Davis, 1991; Chatfield, 2004). With this approach, all entries on the same diagonal of the correlation matrix are the same. For example, with an AR(1) process we need one additional weighting parameter, φ, and $AC_1(l) = \varphi^l$ (see Appendix for correlations for an AR(2) process). Note that this equation describes correlations in **X**, and not in the composition **O**. In particular, when $\varphi > 0$, all correlations in **X** will be positive, but correlations in **O** may be both positive and negative (as will be shown below). The logistic-normal does not allow zero proportions, which is a limitation for stock assessment applications, but not usually, I think, a serious one (see below).

Another likelihood that was designed for continuous composition data (and is thus both proper and correctly self-weighting) is the *Dirichlet*, which was first used for fishery composition data by Williams and Quinn (1998), and has since been used in stock assessments (e.g., Chassot et al., 2009). Its derivation is analogous to that of the logistic-normal in that it involves the application of a transformation which forces the resulting vector to be a composition: a composition **O** is Dirichlet with parameters {**E**, α_0} if $O_b = X_b/\sum_{b'}X_{b'}$, where the X_b are independent gamma variates with shape parameters $\alpha_0 E_b$, and common scale parameter α_0 (which means that the likelihood is given by $\mathrm{NLL} = -Y\log(\Gamma(\alpha_0)) + \Sigma_{by}[\log(\Gamma(\alpha_0 E_{by})) - (\alpha_0 E_{by} - 1)\log(O_{by})]$). The Dirichlet is parsimonious (with the single weighting parameter α_0) but does not allow substantial correlations (it has exactly the same correlation structure as a multinomial, i.e., the correlation between proportions in bins b and c is $\left[-E_b E_c/\left\{(1-E_b)(1-E_c)\right\}\right]^{0.5}$). (Note that the Dirichlet parameters **E** and α_0 used here can be transformed to the more conventional single vector parameter $\boldsymbol{\alpha}$, by setting $\alpha_b = E_b\alpha_0$, which means that $\alpha_0 = \Sigma_b\alpha_b$). Like the logistic-normal, the Dirichlet does not allow zero proportions.

Another likelihood that has sometimes been used in stock assessments is that for the *Dirichlet-multinomial* compound distribution (e.g., Gazey et al., 2008; Hillary, 2011). A composition **O** is Dirichlet-multinomial with parameters {**E**, α_0, N} if it has a multinomial distribution with parameters {**E**′, N}, where **E**′ is Dirichlet with parameters {**E**, α_0}. Like the multinomial, it does not allow substantial correlations, and is a discrete distribution applied to continuous observations. Since its integral over all permissible observations depends on N, but not α_0, it is partially self-weighting (i.e., we can estimate α_0 within the model, as long as we fix N). Thus, in respect of our four desirable properties, it performs better than the multinomial, but not as well as the Dirichlet. Another compound likelihood worth mentioning, though I'm not aware of its being used in a stock assessment, is the logistic-normal-multinomial. This was used by Hrafnkelsson and Stefánsson (2004) in estimating observation error for survey length compositions (though they called it a Gaussian-multinomial and used it for numbers, rather than proportions, at length). In a stock assessment setting this likelihood could allow substantial correlations but, as with the Dirichlet-multinomial, it would be only partially self-weighting.

Table 3
Details of the 72 stock assessment composition data set that were used to evaluate the logistic-normal and Dirichlet likelihoods[a].

Assessment program	Number of assessments	Composition type		
		Age	Length	Likelihood used
Stock Synthesis (Methot and Wetzel, 2013)	6	1	15	Multinomial
CASAL (Bull et al., 2012)	4	11	1	Multinomial
	1	1		Multivariate lognormal
MULTIFAN-CL (Fournier et al., 1998)	3		7	Multivariate normal
BAM (Craig, 2012)	2	3	2	Multinomial
	1	2		Multivariate normal
ASAP (Legault and Restrepo, 1999)	3	12		Multinomial
iSCAM (Martell, 2011)	5	12		Logistic-normal
Ad hoc[b]	3	2	3	Multinomial
	28	44	28	

[a] More information about these data sets is provided in the Supplementary Data.
[b] Programs written specifically for a single species or stock assessment.

Amongst the likelihoods considered here, the logistic-normal is the clear winner in terms of performance with respect to our four desirable properties (Table 2). In the next section this likelihood will be further evaluated using composition data from a wide range of stock assessments. The Dirichlet likelihood will be included in the evaluation because it performs almost as well, and may be appropriate for composition data where correlations are not large. The other likelihoods in Table 2 are excluded because the AIC, used to compare goodness of fit of different likelihoods, would be compromised by the fact that they are not proper.

4. Evaluation of the logistic-normal and Dirichlet likelihoods

For this evaluation I assembled a collection of 72 composition data sets from 28 stock assessments originating from nine different computer programs (Table 3). 44 data sets were for age, and 28 for length. Most assessments used the multinomial likelihood, but there were reasonable numbers from the multivariate normal and logistic-normal, and one multivariate lognormal. The intention was to maximise diversity amongst the data sets, with the hope that any general conclusions from this study would apply to most stock assessments. The greatest weakness in this respect was that only seven data sets were sexed (i.e., the compositions were proportions by age (or length) and sex). In two of the sexed data sets – for New Zealand rock lobster – the concept of sex was extended to three categories: male, immature female, and mature female. Each data set comprised observed and expected proportions, and assumed sample sizes by year (i.e., O_{by}, E_{by}, and N_y). All data sets covered at least 15 y, and most were longer than 20 y, up to a maximum of 124 y (for simplicity I refer to the time steps as years, but in a few assessments the time step was a quarter). Many of the assessments applied some sort of robustification for the compositions, modifying the likelihood (as in MULTIFAN-CL – see Fournier et al., 1998) and/or the data (e.g., Stock Synthesis optionally adds a user-provided small number [10^{-4} by default] to all observed and expected proportions and then renormalizes the data to sum to 1 in each year; in iSCAM, bins with proportions less than a user-specified minimum (default 0.02) are amalgamated with adjacent bins). Data sets in which the data modification affected more than a third of observed proportions were excluded.

Two modifications were made to these data sets. The first was to suppress zeroes (because these are not allowed by either the logistic-normal or Dirichlet likelihoods), which was necessary in 54 data sets. This was done by compressing the tails of the compositions into plus and/or minus groups (i.e., replacing the vector $(O_1,\ldots,O_B)$ by $\left(O'_{lo}, O'_{lo+1}, \ldots, O'_{hi-1}, O'_{hi}\right)$, where $lo \geq 1$, $O'_{lo} = \sum_{b=1}^{lo} O_b$, $hi \leq B$, and $O'_{hi} = \sum_{b=hi}^{B} O_b$, and making an analogous replacement for $\mathbf{E}$). For each data set, the same plus and/or minus groups were used for all years to simplify the analysis of correlation structure. In some data sets this caused an excessive reduction in the number of bins because there were some years (typically with small assumed sample sizes) in which there were many zeroes. To avoid this excessive reduction, years with sample sizes less than some threshold were excluded (the threshold varied between data sets and was set arbitrarily to find what seemed the best trade-off between the number of years and the number of post-zero-suppression bins in the data set). The second modification involved increasing the small constant used to robustify two data sets from each of two assessments. In one assessment, where observed zeroes had been replaced by 10^{-12}, this was replaced by 10^{-5}; in the other, where 10^{-7} had been added to all observed and expected proportions (and then the data were normalised to sum to 1), this was replaced by 10^{-4}. These larger robustifying constants seemed more appropriate because they were closer to the smallest non-zero proportion before robustification (1.5×10^{-5} for the first assessment and 2.5×10^{-4} for the second). In all four data sets, changing the robustifying constant strongly affected the fits to the logistic-normal and Dirichlet likelihoods (see below).

Each data set was fitted to four alternative likelihoods: the Dirichlet, and three versions of the logistic-normal which I denote LN1, LN2, and LN3, with the integer in these labels referring to the number of weighting parameters. LN1 is the logistic-normal used in iSCAM, with the single weighting parameter σ, and LN2 and LN3 are the new extensions described above (see Section 3.1) in which the correlation structure of $\mathbf{X}$ derives from an AR(1) or AR(2) process, respectively (the weighting parameters are $\{\sigma, \varphi\}$ for LN2, and $\{\sigma, \varphi_1, \varphi_2\}$ for LN3). In this fitting (done in R (R Core Team, 2013), using the nonlinear minimization function nlm, and function ARMAacf to calculate the LN3 correlations), only the weighting parameters were estimated (i.e., the likelihood parameters E_{by} were taken to be those estimated in the original assessments). Between-year weighting of the data was achieved by setting $\alpha_{0y} = \alpha_0 \left[N_y/\text{mean}_{y'}\left(N_{y'}\right)\right]$ for the Dirichlet, or $\sigma_y = \sigma\left[\text{mean}_{y'}\left(N_{y'}\right)/N_y\right]^{0.5}$ for the logistic-normal (the correlation parameters for LN2 and LN3 were assumed to be the same for all years). The goodness of fit to the four likelihoods was compared using AIC (Akaike, 1974). Some properties of the fitted distributions (predicted correlations in Section 4.1, and bias in Section 4.3) were investigated using simulated data. For each combination of data set and fitted distribution, I generated 1000 simulated sets of observations of the same size (denoted S_{byi} for $i = 1,\ldots, 1000$), using the associated likelihood and parameters (see R function Simcomp in the Supplementary Data).

Three quarters of the data sets (55/72) preferred (i.e., were better fitted by) one of the correlated logistic-normal likelihoods (LN2

Table 4
Number of data sets, by type (length or age, unsexed or sexed), which preferred (i.e., were better fitted by) each likelihood[a].

	Length data		Age data		
Likelihood	Unsexed	Sexed	Unsexed	Sexed	All
Dirichlet	0	0	14	3	17
LN1	0	0	0	0	0
LN2 or LN3	26	2	25	2	55

[a] More information about the fits to these data sets is provided in the Supplementary Data.

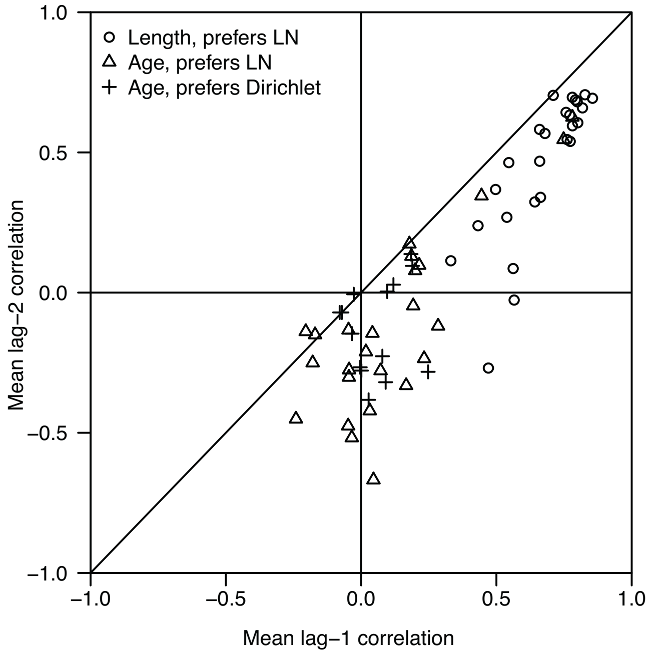

Fig. 2. Mean lag-1 and lag-2 residual correlations for each unsexed data set. Each plotted point represents a single data set, with the plotting symbol identifying data type (length or age) and preference (logistic-normal (LN) or Dirichlet). The mean lag-k residual correlation was calculated as mean$_b$[Cor$_y$($O_{by} - E_{by}$, $O_{b+k,y} - E_{b+k,y}$)].

or LN3), with the remainder preferring the Dirichlet (Table 4) (LN2 and LN3 are grouped in this table because the difference in their AICs was usually small). The preference for the logistic-normal was strongest for the length data (28/28), less strong for the unsexed age data (25/39), and unclear for the sexed age data (2/5). When all the sexed data were converted to unsexed, preference for the logistic-normal increased from 4/7 to 6/7. When the preference for the logistic-normal was quantified as a difference of AICs it was typically much stronger for the length than for the age data (Fig. 1). This was perhaps because correlations were usually stronger in the length data than in the age data (Fig. 2), which is reflected in estimates of the correlation parameter φ (for fits to LN2), which were typically higher for the length data (median 0.81) than the age data (median 0.49). As mentioned above, changing their robustifying constants strongly affected fits to four data sets. When these constants were changed back to the original values the range of AIC differences for the four data sets changed from +22 to +1952 (i.e., all favouring the logistic-normal) to −3123 to −692 (all strongly favouring the Dirichlet). The effect can also be seen in the LN3 estimates of σ, which increased from a range of 1.1–1.9 to 3.4–4.8.

One hypothesis to explain why some age data sets preferred the Dirichlet is that the correlations in these data sets could be small enough to have little effect. To investigate this hypothesis we need a measure of the overall effect of the correlations in a data set. One such measure is the ratio $N_{\text{indiv}}/N_{\text{Penn}}$, where N_{indiv} is an estimate of the effective sample size based on individual residuals (i.e., $O_{by} - E_{by}$), as opposed to the mean age or length residuals which are used to calculate N_{Penn}. A large value of this ratio would suggest that the correlations have a strong effect, whereas a value near 1 would indicate that the effect is small. If the hypothesis were correct then this ratio would typically be higher in age data sets preferring the logistic-normal, and lower in those preferring the

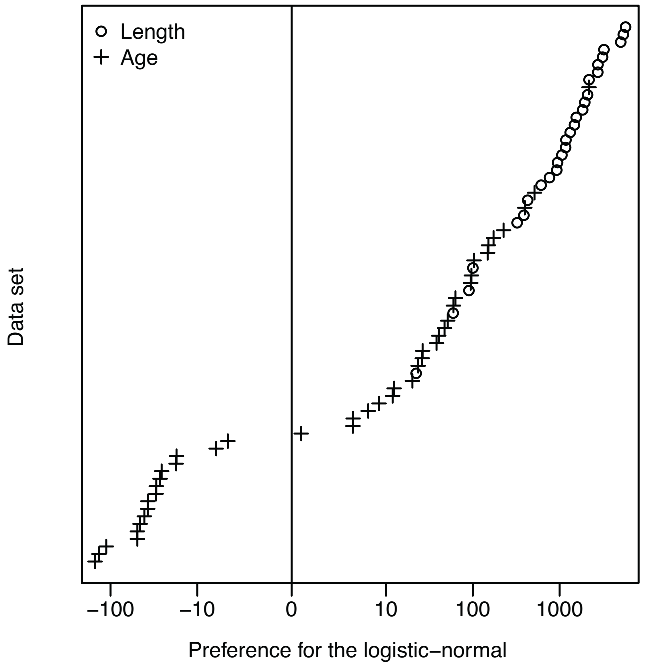

Fig. 1. Strength of the preference of each length or age data set for the logistic-normal likelihood. The preference is represented by a difference of AICs: [AIC$_{\text{Dirichlet}}$ − min(AIC$_{\text{LN1}}$, AIC$_{\text{LN2}}$, AIC$_{\text{LN3}}$)].

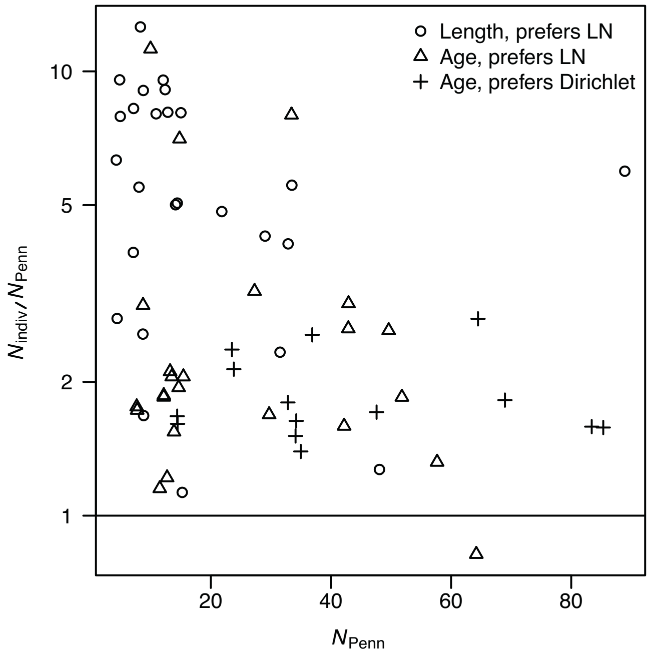

Fig. 3. The ratio $N_{\text{indiv}}/N_{\text{Penn}}$ (a measure of the overall strength of correlations) plotted against N_{Penn} for each unsexed data set. Each plotted point represents a single data set, with the plotting symbol identifying data type (length or age) and preference (logistic-normal (LN) or Dirichlet). One extreme outlier (N_{Penn} = 2.4, $N_{\text{indiv}}/N_{\text{Penn}}$ = 0.026) was omitted to improve plotting clarity.

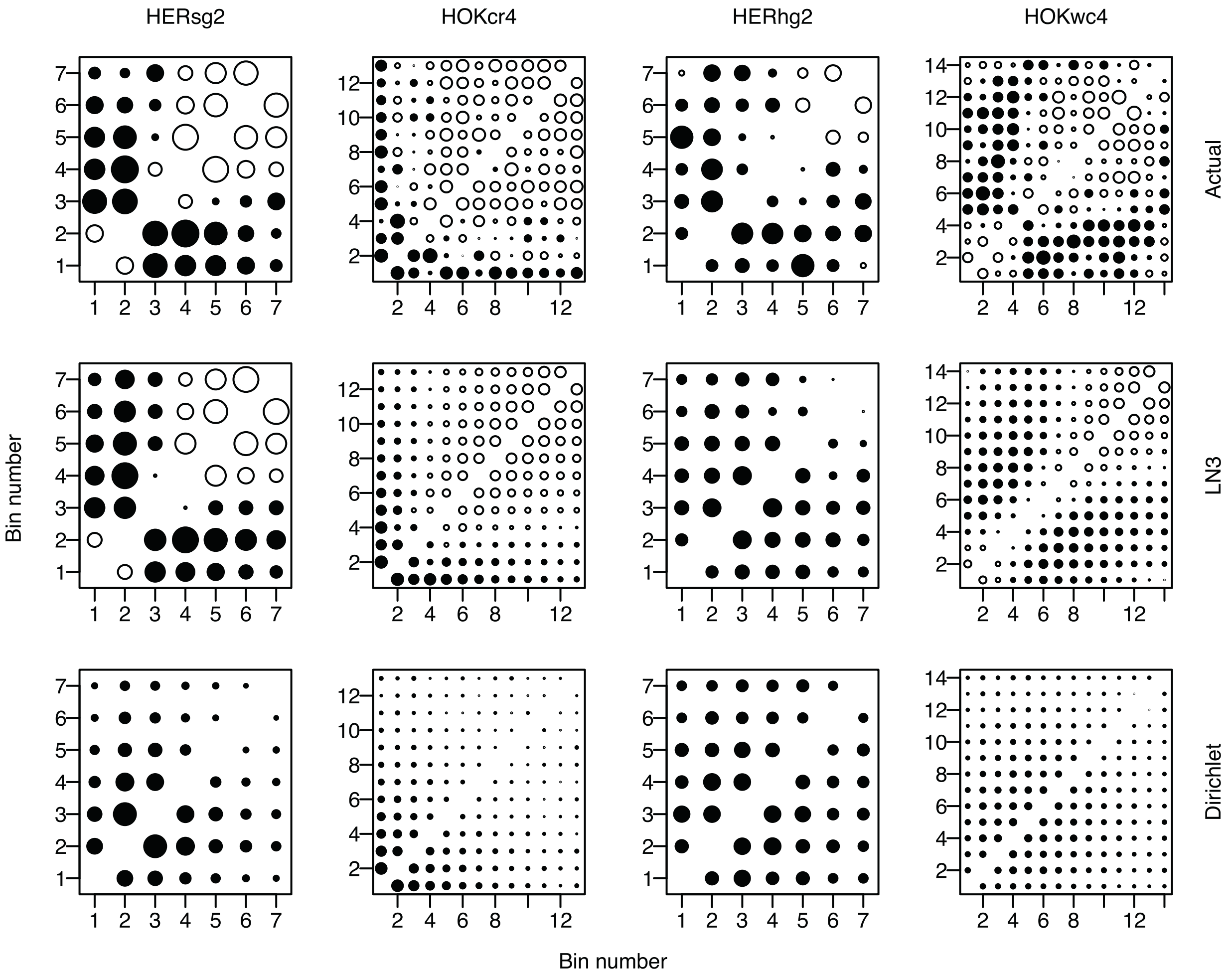

Fig. 4. Comparison of the actual residual correlations (top panels) in four data sets with those predicted by the LN3 (middle panels) or Dirichlet (bottom panels) distributions fitted to them. The correlation between bin numbers *b* and *c* was calculated as the correlation across years of $(O_{by} - E_{by})$ with $(O_{cy} - E_{cy})$. Bubble areas are proportional to the absolute correlations (empty bubbles for positive, filled bubbles for negative). In each panel, bubble sizes are adjusted so that side-by-side bubbles with correlation 1 would just touch. The four data sets were chosen as examples of two where the LN3 replicated the actual correlations reasonably well (HERsg2 and HOKcr4, left panels), and two where it didn't (HERhg2 and HOKwc4, right panels).

Dirichlet. Several ways of calculating N_{indiv} have been proposed. Of three listed by Francis (2011) (methods TA1.1, TA1.2, and TA1.3 in his Table A1), I used the last. The ratio was typically higher for the length data (as expected), but it didn't clearly separate the age data sets that preferred the Dirichlet from those that preferred the logistic-normal, and so did not support the hypothesis (Fig. 3). (The same was true when the plot was repeated using methods TA1.1 or TA1.2 instead of TA1.3).

4.1. How well are correlations replicated?

It is of interest to ask how well the actual correlations in the data were reproduced by the fitted distributions. A visual evaluation of plots comparing actual and predicted correlations showed that the logistic-normal sometimes performed well, and sometimes less so, whereas the Dirichlet performance was usually poor, because of its inability to produce positive correlations (Fig. 4) (the predicted correlations for LN3 were estimated from the simulated data because there is no closed form equation for these correlations; the same procedure was used for the Dirichlet for reasons of comparability). The logistic-normal was preferred for three of the four data sets in Fig. 4: all but HERsg2.

Of course we should not expect to get a perfect match between actual and predicted correlations in Fig. 4 because the expected proportions (E_{by}) are not those that would be produced had the likelihoods been fitted within the stock assessment model. Also, the plotted 'actual' correlations are only approximate because the sizes of the samples from which they were calculated (the number of years in each data set) are quite small (41 for the HER data; 21 and 24 for the HOK data). For example, with uncorrelated normal data the standard error in correlations estimated from samples is about 0.16 for $n = 41$, and 0.22 for $n = 21$. Further, it is very likely that the true correlations will vary from year to year. This certainly seems to be the case for the observation-error correlations from the HOKcr4 data set (Fig. 5).

I looked more closely at the comparison between actual and predicted correlations for the 12 herring data sets that came from assessments using the LN1 likelihood. Concentrating on the four data sets that most strongly preferred the Dirichlet likelihood, I found that the consistency between predicted and actual correlations was usually greater, and never less, for the LN3 than for the Dirichlet (Fig. 6).

4.2. The problem of sex

Sex is a substantial complication when seeking an appropriate correlation structure for a composition likelihood. The simple autoregressive correlation structure used in the LN2 and LN3 likelihoods is not adequate for sexed data because it depends on bin numbers, and so is affected by the order of the bins. For an unsexed

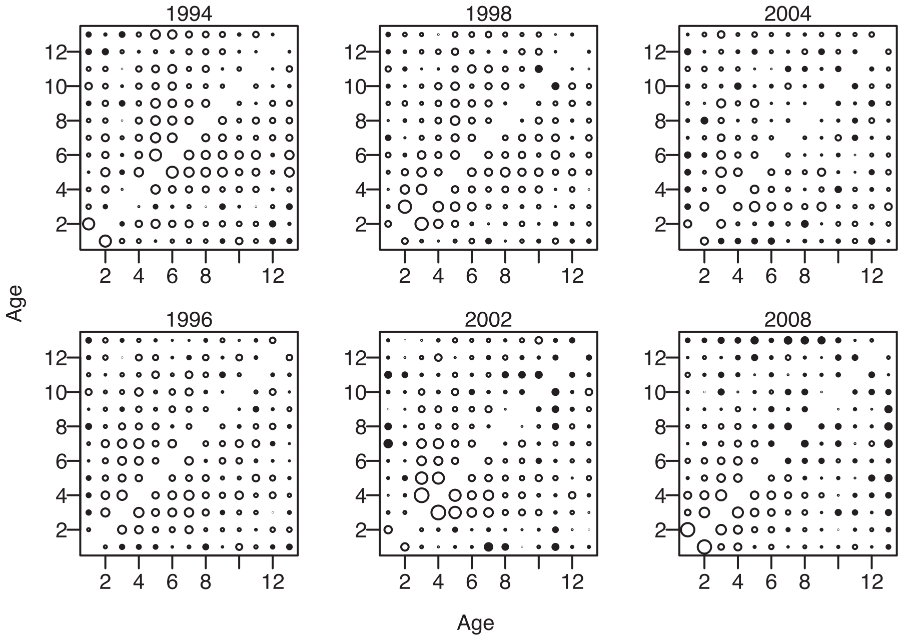

Fig. 5. Observation-error correlations (estimated by bootstrap resampling [$n = 300$] of the raw data) for selected years from the HOKcr4 data set (from a series of surveys). Bubble plotting conventions as in Fig. 4.

composition this produces the same correlation structure whether the bins are ordered by increasing or decreasing age (or length). However, different plausible ordering of the bins for sexed data produce different correlation structures. For the sexed data sets in this study the bins were ordered by sex (males then females or, for the rock lobster data, males, immature females, then mature females), and within sex by increasing age or length. Changing the order of the sexes changes the autoregressive correlation structure, as does deciding to order first by length or age, and within that by sex. Also, the order used in this study leads to an asymmetry: for example, the correlation between male 3-year olds and female 5-year olds differs from that between male 5-year olds and female 3-year olds.

The difficulty of finding an appropriate correlation structure for sexed data is underlined by the fact that some exploratory plots found striking heterogeneity amongst our very small sample of five sexed age data sets. For example, looking at between-sex correlations, those for lag-0 and lag-1 were very different for three data sets (those for species HOK), but similar for the others (Fig. 7). Another plot was intended to explore the possibility that correlations might depend only on the difference in ages. This would mean that within- and between-sex correlations for the same pair of ages would be similar, which seemed to be true for two of the data sets (HOKcr3 and HOKwc3), but not the other three (Fig. 8).

There are other difficulties. The correlations we are interested in (e.g., Figs. 7 and 8) are of the observations **O** (or, strictly speaking, of the residuals, **O** − **E**), but the autoregressive correlations specified for LN2 or LN3 apply to the multivariate normal **X**, which is logistically transformed to **O**. Another difficulty is that correlation structures that appear conceptually appealing may not always be statistically valid. Consider, for example, the structure mentioned in the previous paragraph, which we might implement by setting the correlation (for **X**) between bins for ages a and b equal to $AC_2(|a-b|+1)$, independent of the sex associated with these bins (we need the '+1' here to allow for between-sex correlations when $a = b$). It turns out that there are pairs of parameters φ_1, φ_2, which are valid for an AR(2) process but which produce a matrix of correlations which is not statistically valid (i.e., not positive definite).

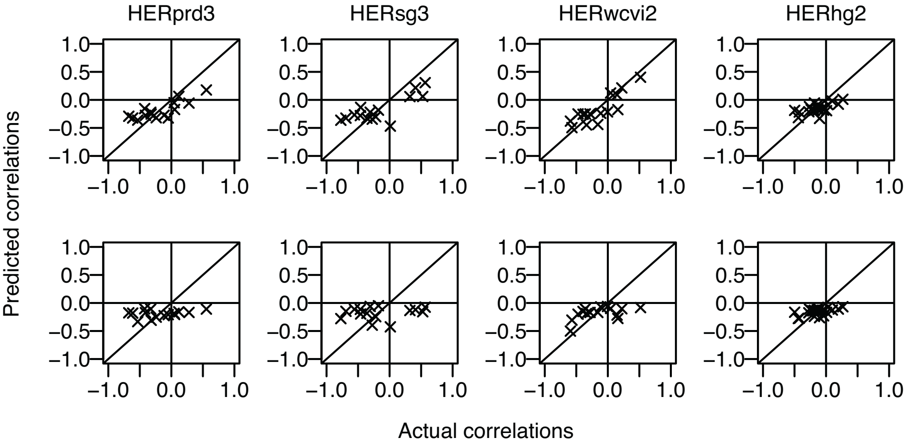

Fig. 6. Actual correlations (x-axis) plotted against those predicted from fitted distributions (y-axis) using the LN3 (upper panels) or Dirichlet (lower panels) distributions for the four of the 12 herring data sets (which came from an assessment using the LN1 likelihood) which most strongly preferred the Dirichlet.

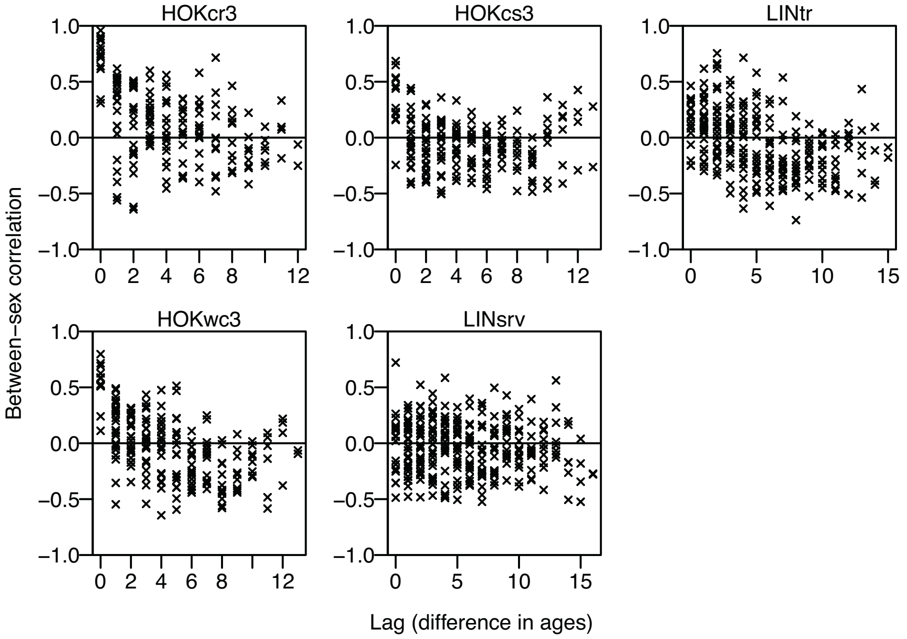

Fig. 7. Between-sex residual correlations plotted against lag (difference in ages) for the five sexed age composition data sets. For each plotted point the y-value is the correlation between the proportions of males at some age a and the proportions of females at some age b and the x-value is $|a - b|$. The data sets are from the sexed model of the hoki assessment presented in Table 1 and the 2011 New Zealand assessment of Chatham Rise ling (*Genypterus blacodes*) (tr = trawl fishery, srv = trawl survey).

Further, whether a particular pair of parameter values will produce a valid correlation matrix depends on the number of age (or length) bins.

4.3. Two types of bias

The logistic-normal differs from the Dirichlet (and multinomial) in an important way that seems, at least at first, a great disadvantage. For the Dirichlet with parameters (**E**, α_0), the statistical expectation of an observation in bin b is E_b, but this is not true for a logistic-normal with parameters (**E**, σ) (although it is approximately true for small σ, as Schnute and Haigh (2007) noted). Since there is no closed form equation for the expected values from a logistic-normal (Aitchison, 2003) I used the simulated data to illustrate how big a problem this property of the logistic-normal might be. For this I chose four unsexed data sets whose estimated values of σ (for LN1) covered a wide range: two age data sets with lower values – HERprd3 (σ = 0.43) and GMCaind1 (0.67) – and two length data sets with higher values – PAUcs (0.81) and SKJ5 (1.03). The ratios, mean$_i(S_{byi})/E_{by}$, were taken as estimates of the degree of

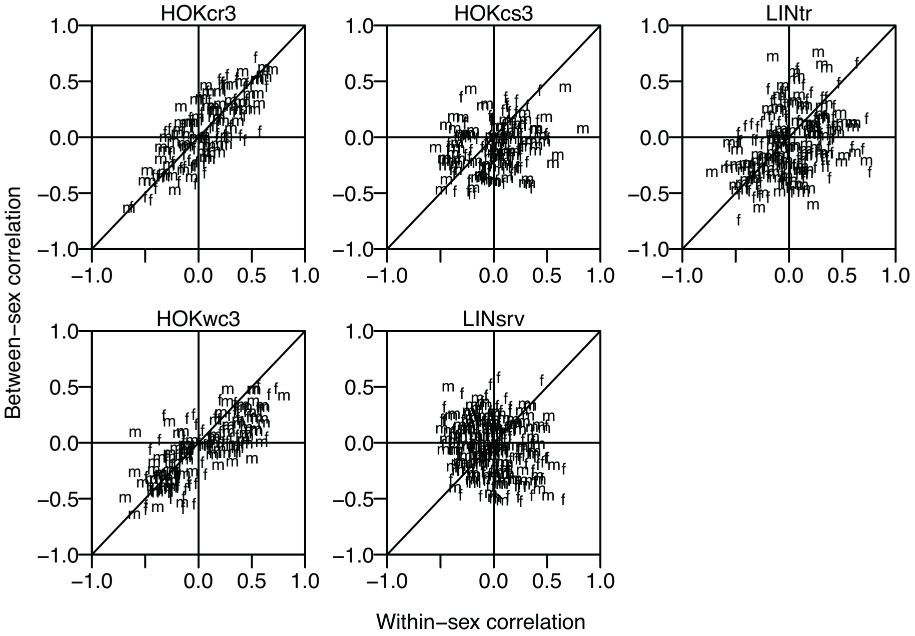

Fig. 8. Comparison, for the five sexed age composition data sets of Fig. 7, of within- and between-sex residual correlations for the same pair of ages. For each ordered pair of distinct ages, a and b, two points are plotted: $C_{Ma,Mb}$ (x-axis) against $C_{Ma,Fb}$ (y-axis) (plotting symbol 'm'); and $C_{Fa,Fb}$ (x-axis) against $C_{Fa,Mb}$ (y-axis) (plotting symbol 'f'), where $C_{Xa,Yb}$ denotes the correlation between the residuals of proportions at sex X and age a and those at sex Y and age b.

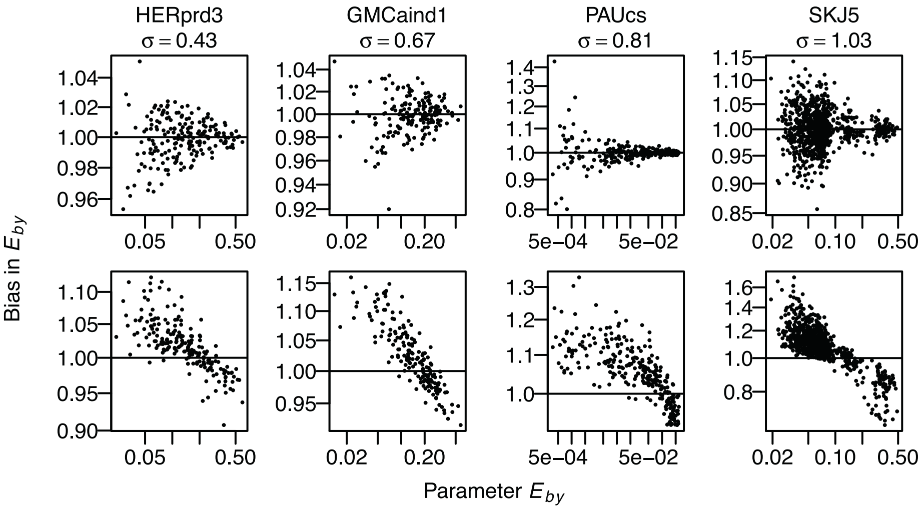

Fig. 9. Bias, estimated from simulated data based on four data sets (HERprd3, GMCaind1, PAUcs, and SKJ5), associated with the parameter E_{by} for Dirichlet (upper panels) and LN1 (lower panels) likelihoods. The y-axis represents mean$_i$(S_{byi})/E_{by}, so an even scatter of points about the horizontal y = 1 line indicates no bias. The four data sets are ordered by increasing σ (as estimated in the fit to the LN1 likelihood) – see values above plot.

bias associated with the expected value (values close to 1 indicate little bias), and these ratios were calculated from the simulated data sets associated with both the LN1 and Dirichlet distributions. Plots of the ratios against E_{by} show no bias for the Dirichlet (note that the points are scattered evenly about the y = 1 line), but a bias for the LN1 distribution which was greater in the data sets with higher σ (Fig. 9). For LN1, simulated values tended to be higher than E_{by} when E_{by} was small, and lower when it was large.

Aitchison (2003) maintained that this bias in the logistic-normal is of no concern. He suggested that, because composition data are only relative (having to sum to 1), what is important is their ratios (O_b/O_c), rather than their actual values (O_b). This makes sense in a stock assessment context. In terms of the quantities we wish to estimate from compositions (mainly mortality rates, relative year-class strengths, and selectivity) we can learn very little by simply observing the catch proportion at age 5, say, in each year (i.e., O_{5y}), whereas the ratios O_{5y}/O_{6y} are much more informative about these quantities of interest. Moreover, because ratios of composition proportions typically cover a very wide range of values (several orders of magnitude would be common for fishery compositions) Aitchison suggested it is sensible to consider them in log space. In this domain, the logistic-normal is unbiased (i.e., the statistical expectation of log(O_b/O_c) is equal to log(E_b/E_c), as proved by Aitchison), but the Dirichlet is biased. To demonstrate this I calculated, from the same simulated data as used for Fig. 9, the difference mean$_i$[log(S_{byi}/S_{cyi})] – log(E_{by}/E_{cy}) as a measure of bias. For the LN1 data this measure was scattered evenly around 0, indicating no bias, whereas for the Dirichlet it showed a strong increasing trend as the ratio E_{by}/E_{cy} increased (Fig. 10).

To understand how these biases might affect output from an assessment we need to remember that, compared to this simulation experiment, the order of things is reversed in an assessment. Here, we start with known values of E_{by} and generate simulated observations; in an assessment we start with known observations and estimate the E_{by}. What the present results suggest for stock assessments is that, for the logistic-normal, the estimated E_{by} will tend to be more extreme than the observations, O_{by} (e.g., values of b and y for which O_{by} is small will tend to have estimates, E_{by}, that are even smaller). This will not be true for the Dirichlet, but for this distribution the estimated ratios E_{by}/E_{cy} will tend to be less

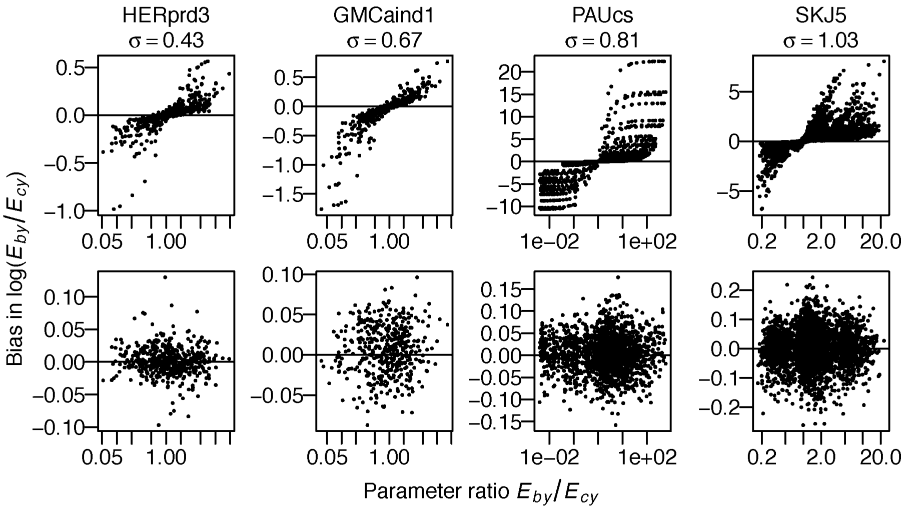

Fig. 10. Bias, estimated from simulated data based on four data sets (HERprd3, GMCaind1, PAUcs, and SKJ5), associated with log(E_{by}/E_{cy}) for Dirichlet (upper panels) and LN1 (lower panels) likelihoods. The y-axis represents mean$_i$[log(S_{byi}/S_{cyi})] – log(E_{by}/E_{cy}), so an even scatter of points about the horizontal y = 0 line indicates no bias. The four data sets are ordered by increasing σ (as estimated in the fit to the LN1 likelihood) – see values above plot.

extreme than the corresponding observed ratios, O_{by}/O_{cy}, so the range of estimated year-class strengths will tend to be narrower than is consistent with the observation (but this will not be true for the logistic-normal). Thus, in choosing between these two likelihoods we need to decide which is the more important type of bias. Aitchison (2003) argued that the latter bias matters more, so the logistic-normal is preferable (at least with regard to bias). Of course, things are not quite as clear cut at these results might suggest. Our conclusion that the logistic-normal likelihood will not produce bias in $\log(E_{by}/E_{cy})$ (and that the Dirichlet will not produce bias in E_{by}) is conditional on the composition error distribution being exactly as assumed by the likelihood.

5. Conclusions/Discussion

I hope I have shown that there is a need to replace the multinomial as the likelihood of choice for composition data in stock assessment models. Though it has the advantage of simplicity, it is simply wrong for the task at hand. When used in a stock assessment model it is improper (because it is a discrete likelihood being used for continuous data), and thus can't be self-weighting, and it fails to mimic the correlations common in composition data. Use of clearly inappropriate likelihoods compromises statistical inferences from the assessment model (e.g., confidence intervals, or AIC for model selection). Two common alternatives to the multinomial – the multivariate normal and lognormal – are not much better. They could possibly be made to produce appropriate correlation structures, but because they ignore the defining property of compositions (that they are sets of non-negative numbers summing to 1), they are also improper, and to use them for these data is like using a chisel to drive a screw. This is something we should do only when there is no more appropriate tool available. Aitchison (2003), who complained about the misuse of standard statistical techniques with composition data, has provided a tool – the logistic-normal – that shows great potential for this task, at least for unsexed compositions. It is designed specifically for continuous composition data (unlike the other likelihoods just mentioned), and so is proper and self-weighting, and seems to be able to produce appropriate correlation structures with few parameters (at least for unsexed data).

The only other likelihood in Table 2 that offers any promise is the Dirichlet. Like the logistic-normal, it was designed for composition data, so it is proper and self-weighting. Its great weakness is its inflexibility with regard to correlations. It may find a use for composition data sets where correlations are small, but it would be much more useful if it could be generalized so as to incorporate additional parameters which allowed substantial correlations. Unfortunately such generalizations do not yet seem to be available (see chapter 13 of Aitchison, 2003).

The analyses presented above are only the first step in evaluating the logistic-normal for use in stock assessments. Their power is that they were applied to a wide range of length and age composition data from many assessments and computer programs. Thus we might expect that conclusions from these analyses would be applicable to most assessments and data sets. However, their weakness is that they were applied after the assessments, rather than as a part of them. The fits we obtained, estimating only the weighting parameters, may be very different from those that would have been obtained had we done the fitting within the various assessment models (but that, of course, would have been a very substantial task, involving redoing 28 stock assessments in nine different computer programs).

The next step will be to code the logistic-normal likelihood into stock assessment programs and see how it performs. In the hope of encouraging other researchers to join me in this task I close by discussing some of the technical matters associated with it (with more details in the Appendix).

5.1. *Using the logistic-normal in stock assessments*

There are three matters that need to be addressed to enable a thorough evaluation of the logistic-normal as a composition likelihood in stock assessments models: zeroes, robustification, and sexed data.

There are many ways of dealing with zeroes in composition data, but perhaps the simplest would be a judicious mixture of tail compression (see Section 4) and replacement by a small number, ε (after which, each composition should be normalized to sum to 1). If we think of some zeroes as arising by *happenstance* (i.e., these observations could, by chance, have been non-zero), and others as being *definitive* (i.e., they would never be non-zero), then our aim would be to remove definitive zeroes by tail compression, and replace happenstance zeroes by ε. Both these techniques (or something similar) are already commonly used (as a form of robustification) with the multinomial likelihood. The tail compression can (and probably should) be done independently for each year in the data set (this was not done above so as to allow simpler examination of correlation structures). It doesn't seem sensible to replace more than a small percentage (perhaps 5% or less) of zeroes by ε (this percentage exceeded 50% in several data sets considered for, but excluded from, the above analyses). A reasonable choice for ε would be something a bit smaller than the smallest remaining non-zero proportion; when zeroes arise from rounding, then it would seem sensible to set ε to the maximum rounding error (e.g., when rounding proportions to three decimal places set $\varepsilon = 0.0005$). Another method of removing zeroes is to amalgamate bins with zero proportions with adjacent bins. This seems reasonable when the zero bins are randomly scattered through the composition data, but could be problematical when these zeroes are associated with one or more exceptionally weak year classes (Martell, 2011).

The choice of a value for the above small number, ε, is sometimes rather subjective and so it is often advisable to test how robust an assessment is to changes in this value. Aitchison (2003) suggested a more objective approach for the situation in which the zeroes may be thought of arising from rounding: replace each zero by $\varepsilon = \delta(n_0+1)(B-n_0)/B^2$ and subtract $\delta n_0(n_0+1)/B^2$ from each non-zero value, where n_0 is the number of zeroes in the composition and δ is the maximum rounding error (e.g., 0.0005 if proportions are rounded to three decimal places). The logic behind this is that if we think of each composition as a point in B-dimensional space, then rounding generates a region of uncertainty about that point, and Aitchison's procedure replaces each zero-containing composition with the point at the geometric centre of its region of uncertainty. Note that with this procedure ε will not generally be the same for each composition in a data set (because n_0 will usually vary between compositions).

Rather than removing zeroes, we could adjust the likelihood to allow them. Aitchison (2003) suggested assuming that $\mathbf{O}+\tau$ is logistic-normal, where τ is a small number which may be fixed or estimated [this is analogous to a common method of allowing for zeroes in a univariate lognormal distribution (Aitchison and Brown, 1957)]. With this assumption it would be sensible to similarly adjust the role of the parameter $\mathbf{E}$ [i.e., assume that the multivariate normal $\mathbf{X}$, for which $\mathbf{O}+\tau=\exp(\mathbf{X})/\Sigma_b\exp(X_b)$, has mean $\log(\mathbf{E}+\tau)$]. More complicated adjustments are possible. For example we could assume that each O_b has independent probability $f(E_b)$ of being zero (where f is some simple increasing function whose parameters could be estimated) and the distribution of the non-zero part of $\mathbf{O}$ is logistic-normal. Though such models may sometimes be called for I suspect that zeroes in compositions are usually more sensibly treated as a nuisance to be obviated, rather than a phenomenon to be modelled.

There are two factors which together make it important to robustify composition likelihoods. First, composition data include many more individual data points than do abundance data (the other common data type in stock assessments). This means that the signal in the latter data can easily be swamped by that in the former. Second, the real world is much more heterogeneous than the simplistic world of our stock assessment models. This means that it is common to find, amongst the large number of individual composition data points, a few that are strong outliers (i.e., they appear highly improbable compared to the majority of data points), as noted by Fournier et al. (1990). Thus there is a need to modify composition likelihoods to make them robust to these outliers (i.e., to reduce the influence of these outliers on the fit of the model to other data). (Such modification may make the likelihood theoretically improper, and thus (probably slightly) bias estimates of weighting parameters, but this disadvantage will usually be strongly outweighed by the gain from reducing sensitivity to outliers.) How we should modify a likelihood to make it more robust depends strongly on its mathematical form. I avoided such modifications in the above evaluation of the logistic-normal and Dirichlet likelihoods because it was not possible to do this in a balanced way (i.e., to achieve the same level of robustification in both likelihoods). How best to robustify the logistic-normal remains an open question. It might seem straightforward to adapt the approach of Fournier et al. (1998), but my attempt to do this was not successful (see Appendix).

Perhaps the biggest challenge for the logistic-normal is to deal with sexed compositions. Note that this challenge is not restricted to the logistic-normal; it applies to any likelihood as soon as we try to introduce realistic correlations. Two approaches are possible. The first is to find a two-dimensional way of introducing correlations (as opposed to the simple one-dimensional autoregressive approach used above). As noted above, the sexed data sets considered in this study were too few to suggest an obvious solution. A second approach would be to separate sexed compositions into two data sets – one for the age or length structure, and the other for the sex structure – and to provide a likelihood for each. This requires the assumption that the sampling error for the two data sets be statistically independent, which is probably untrue, but may be an acceptable approximation to the truth.

Finally, researchers are urged to take advantage of the flexibility of the logistic-normal. If the forms given above (LN1, LN2, LN3) are inadequate for your data, then consider devising new forms (see, e.g., LN3m in the Appendix). In all forms of the logistic-normal considered above, all components of the multivariate normal vector **X** have the same standard deviation, σ, but this assumption might usefully be relaxed for some data sets (e.g., by making σ a linear function of bin number). I have assumed that between-year weighting is defined very simply by setting $\sigma_y = \sigma\left[\text{mean}_{y'}\left(N_{y'}\right)/N_y\right]^{0.5}$, but other approaches are possible (e.g., $\sigma_y = \left[\sigma_1^2\left(\text{mean}_{y'}\left(N_{y'}\right)/N_y\right)+\sigma_2^2\right]^{0.5}$, where σ_1 and σ_2 are estimable parameters relating to observation and process error, respectively). Should it be considered desirable to investigate the effect of down-weighting logistic-normal composition data (e.g., to see whether this might allow better fits to abundance data, following Francis, 2011) this can be achieved simply by multiplying σ (or σ_1 and σ_2 for the extension just discussed) by 2 (or some other number greater than 1). Note, however, that the weight assigned to a LN2 or LN3 data set depends on both σ and the other weighting parameter(s) (so, e.g., two LN2 likelihoods with (σ = 0.5, φ = 0.8) and (σ = 0.5, φ = 0.9) imply different data weights).

Acknowledgements

I am grateful to Jon Schnute and Rowan Haigh for their pellucid exposition of the logistic-normal; to J. Aitchison for his extensive development of this distribution in particular, and of composition-analytical methods in general; to those who provided data sets for this study (Cristian Canales, John Field, Dan Fu, Vivian Haist, Peter Horn, Simon Hoyle, Chris Legault, Mark Maunder, Andy McKenzie, Sally Wayte, and Erik Williams), and to André Punt, Steve Martell, and Mark Maunder, for useful comments on a draft of this paper. Permission to use data sets from New Zealand and Australia was granted by the New Zealand Ministry of Primary Industries and the Australian Fisheries Management Authority, respectively.

Appendix A. Technical details of the logistic-normal distribution

In this appendix I provide some technical details that may be useful in implementing the logistic-normal likelihood in stock assessment programs.

In addition to the notation defined above I will denote the arithmetic and geometric means of a vector **X** by $\bar{X}$ and $\tilde{X}$, respectively (so $\bar{X} = \left(1/B\right)\Sigma_{\mathrm{b}}X_b$ and $\tilde{X} = [\Pi_b X_b]^{1/B}$). The symbols Σ_b and Π_b denote sums and products over the range b = 1,. . .,B; whereas Σ_b^- and Π_b^- are for the range b = 1, . . .,(B – 1). $\mathbf{I}_B$ is the $B \times B$ identity matrix, and $\mathbf{J}_B$ is the $B \times B$ matrix with all entries 1.

A.1. Three characterizations of the logistic-normal

There are three related ways to approach the logistic-normal. In the approach used by Schnute and Haigh (2007), but generalised to allow for correlations, a composition **O** has a logistic-normal distribution with parameters (**E**, **C**), if

$$O_b = \frac{\exp\left(X_b\right)}{\sum_{b'}\exp\left(X_{b'}\right)} \text{for} \quad b = 1, \ldots, B \tag{A1}$$

where **X** is multivariate normal with mean log(**E**) and covariance matrix **C**. Note that **O**, **E**, and **X** are all vectors of length B and **C** is a $B \times B$ matrix (in the simple form used by Schnute and Haigh [which I have called LN1], $\mathbf{C} = \sigma^2\mathbf{I}_B$).

Aitchison (2003) used a different approach: starting with a multivariate normal vector **Y** of length (B – 1) with mean $\boldsymbol{\mu}$ and covariance matrix **V**, he defined

$$O_b = \begin{cases} \exp\left(Y_b\right)/\left[1+\sum_{b'}^{-}\exp\left(Y_{b'}\right)\right] & \text{for} \quad b = 1, \ldots, B\text{–}1 \\ 1-\sum_{b'}^{-}O_{b'} & \text{for} \quad b = B \end{cases} \tag{A2}$$

We can derive Eq. (A2) from Eq. (A1) by setting $Y_b = X_b - X_B$ for b = 1,. . .,B – 1, and then replacing X_b in Eq. (A1) by $Y_b + X_B$ (note that the term exp(X_B) cancels from numerator and denominator). We can also express $\boldsymbol{\mu}$ and **V** in terms of **E** and **C**: $\mu_b = \log(E_b/E_B)$ and

$$\mathbf{V} = \mathbf{K}\,\mathbf{C}\,\mathbf{K}^{\mathrm{T}} \tag{A3}$$

where **K** is the (B – 1) × B matrix formed by adding a vector of –1s to the right side of $\mathbf{I}_{B-1}$, and $^{\mathrm{T}}$ denotes the matrix transpose.

An advantage of Eq. (A2) is that it is a one-to-one transformation, unlike Eq. (A1), and so has an inverse transformation

$$Y_b = \log\left(\frac{O_b}{O_B}\right) \tag{A4}$$

The third way to characterise our logistic-normal composition **O** uses the multivariate normal vector **Z**, which is of length B and defined by $\mathbf{Z} = \mathbf{X} - \bar{X}$. If we replace X_b in Eq. (A1) by $Z_b + \bar{X}$ we find

$$O_b = \frac{\exp(Z_b)}{\sum_{b'} \exp(Z_{b'})} \quad \text{for} \quad b = 1, \ldots, B \tag{A5}$$

By taking the product over b on both sides of Eq. (A5) we can show that $\tilde{O} = 1/\sum_{b'} \exp(Z_{b'})$ (using the fact that $\Sigma_b Z_b = 1$), and substituting this back into Eq. (A5) and solving for Z_b produces the inverse transformation

$$Z_b = \log\left(\frac{O_b}{\tilde{O}}\right) \tag{A6}$$

Z has mean log(E/$\tilde{E}$), and Aitchison (2003) has shown that its covariance matrix **Γ** may be calculated as

$$\boldsymbol{\Gamma} = \mathbf{F}^{\mathrm{T}}\mathbf{H}^{-1}\mathbf{V}\mathbf{H}^{-1}\mathbf{F} \tag{A7}$$

where **F** is the $(B-1) \times B$ matrix formed by adding a vector of 1s to the right side of $\mathbf{I}_{B-1}$, and $\mathbf{H} = \mathbf{I}_{B-1} + \mathbf{J}_{B-1}$.

The three characterizations of the logistic-normal (using **X**, **Y**, and **Z**) are useful in different settings. For stock assessments, the Schnute and Haigh (2007) approach (using **X**) is convenient for its parameterisation (because we want the distribution of our B-dimensional composition, **O**, to have amongst its parameters the B-dimensional vector **E** of expected values, rather than the $(B-1)$-dimensional $\boldsymbol{\mu}$). However, it is convenient to use **Y** for the likelihood, and **Z** for standardised residuals.

A.2. The likelihood

The logistic-normal negative log-likelihood may be written as

$$\mathrm{NLL} = 0.5(B-1)\log(2\pi) + \sum\nolimits_b \log(O_b) + 0.5\log\left|\mathbf{V}\right| + 0.5\mathbf{w}^T\mathbf{V}^{-1}\mathbf{w} \tag{A8}$$

where the vector **w**, of length $(B-1)$, is defined by $w_b = Y_b - \mu_b = \log(O_b/O_B) - \log(E_b/E_B)$. This form of the likelihood is taken from equation (C.6) of Schnute and Haigh (2007), who went on to show that, for the simple case of the LN1 distribution, $|\mathbf{V}| = B\sigma^{2(B-1)}$ and $\mathbf{V}^{-1} = \sigma^{-2}[\mathbf{I}_{B-1} - (1/B)\mathbf{J}_{B-1}]$ (see their equations (C.10) and (C.11)).

So far we have considered only a single composition (i.e., one year's data). In fitting the distributions LN1–LN3 to multi-year data we allowed σ to depend on year, defining $\sigma_y = \sigma W_y$, where $W_y = \left[\mathrm{mean}_{y'}\left(N_{y'}\right)/N_y\right]^{0.5}$, but made the other weighting parameters (φ for LN2; φ_1 and φ_2 for LN3) independent of year. For the more general logistic-normal distribution, in which **C** can be any valid covariance matrix, this is equivalent to defining $\mathbf{C}_y = W_y^2\mathbf{C}$, which means that $\mathbf{V}_y = W_y^2\mathbf{V}$. With this assumption it is straightforward to show that the negative log-likelihood for a multi-year composition data set becomes

$$\begin{aligned}\mathrm{NLL} = {} & 0.5Y(B-1)\log(2\pi) + \sum\nolimits_{by}\log\left(O_{by}\right) + 0.5Y\log\left|\mathbf{V}\right| + \\ & (B-1)\sum\nolimits_y \log\left(W_y\right) + 0.5\sum\nolimits_y \frac{(\mathrm{w}_{\mathrm{y}}^{\mathrm{T}}\mathrm{V}^{-1}\mathrm{w}_{\mathrm{y}})}{W_y^2}\end{aligned} \tag{A9}$$

where the vector $\mathbf{w}_y$, of length $(B-1)$, is defined by $w_{by} = Y_{by} - \mu_{by} = \log(O_{by}/O_{By}) - \log(E_{by}/E_{By})$. To evaluate this likelihood for any of LN1, LN2, or LN3, we first construct the covariance matrix **C** from the weighting parameters, calculate **V** from **C** using equation Eq. (A3), and then calculate $|\mathbf{V}|$ and $\mathbf{V}^{-1}$.

In special cases, like LN1–LN3, where we can write C = σ2Č, where Č is the correlation matrix defined by the other weighting parameters (φ for LN2; φ_1 and φ_2 for LN3), it may be useful to rewrite Eq. (A9) as

$$\begin{aligned}\mathrm{NLL} = {} & 0.5Y(B-1)\log(2\pi) + \sum\nolimits_{by}\log\left(O_{by}\right) + (B-1)Y\log(\sigma) + \\ & 0.5Y\log\left|\check{\mathbf{V}}\right| + (B-1)\sum\nolimits_y \log\left(W_y\right) + 0.5\sigma^{-2}\sum\nolimits_y \frac{\left(\mathbf{w}_y^T\check{\mathbf{V}}^{-1}\mathbf{w}_y\right)}{W_y^2}\end{aligned} \tag{A10}$$

where $\check{\mathbf{V}}$= K$\check{\mathbf{C}}\mathbf{K}^{\mathbf{T}}$, because this allows us to treat σ as a nuisance parameter in the stock assessment (as catchability parameters often are) which can be estimated directly (given values for all other parameters) as

$$\hat{\sigma} = \left[\frac{\sum_y \left(\mathbf{w}_y^T\check{\mathbf{V}}^{-1}\mathbf{w}_y\right)/W_y^2}{(B-1)Y}\right]^{0.5} \tag{A11}$$

So far I have assumed that all compositions have the same number of bins (B), which may not be the case if tail compression to suppress zeroes is applied separately by year. For that eventuality, we need to add a subscript y to B and the various matrices, and change Eqs. (A10) and (A11) to

$$\begin{aligned}\mathrm{NLL} = {} & 0.5\log(2\pi)\sum\nolimits_y\left(B_y - 1\right) + \sum\nolimits_{by}\log\left(O_{by}\right) + \log(\sigma)\sum\nolimits_y\left(B_y - 1\right) + \\ & 0.5\sum\nolimits_y \log\left|\check{\mathrm{V}}_{\mathrm{y}}\right| + \sum\nolimits_y (B_y - 1)\log(W_y) + 0.5\sigma^{-2}\sum\nolimits_{\mathrm{y}} \mathrm{w}_y^T\check{\mathrm{V}}_y^{-1}\mathrm{w}_{\mathrm{y}}/W_y^2\end{aligned} \tag{A12}$$

and

$$\hat{\sigma} = \left[\frac{\sum_{\mathrm{y}} \mathrm{w}_y^T\check{\mathrm{V}}_y^{-1}\mathrm{w}_{\mathrm{y}}/W_y^2}{\sum_y\left(B_y - 1\right)}\right]^{0.5} \tag{A13}$$

A.3. Standardised residuals

A useful diagnostic in evaluating the fit of any likelihood to a composition data set is to plot standardised residuals (i.e., residuals which will have mean 0 and standard deviation 1 if the likelihood correctly describes the error structure of the data). Since $\mathbf{w}_y$ is normally distributed with mean 0 and standard deviation $W_yV_{bb}^{0.5}$, an obvious form for standardised residuals for a logistic-normal likelihood is

$$s_{by} = \frac{w_{by}}{W_yV_{bb}^{0.5}} = \frac{\log\left(O_{by}/O_{By}\right) - \log\left(E_{by}/E_{By}\right)}{W_yV_{bb}^{0.5}} \tag{A14}$$

This form has two disadvantages: it is asymmetric (all ratios being formed with the last bin), and it does not provide a residual for every data point (because Eq. (A14) is not defined for $b = B$). Both these disadvantages are avoided if we use a similar approach, but based on **Z** rather than **Y**:

$$s_{by} = \frac{Z_{by} - E\left(Z_{by}\right)}{W_y\Gamma_{bb}^{0.5}} = \frac{\log\left(O_{by}/\tilde{O}\right) - \log\left(E_{by}/\tilde{E}\right)}{W_y\Gamma_{bb}^{0.5}} \tag{A15}$$

(both types of standardised residuals are calculated by R function Sres.logistnorm in the Supplementary Data; the latter, which is recommended, are calculated by default; the former are calculated if argument centred = F).

A.4. Robustifying the logistic-normal likelihood

To illustrate the difficulty of robustifying the logistic-normal I will describe an apparently logical approach (adapted from that of Fournier et al. (1998) for the multivariate normal) which, nonetheless, does not work (with it, the estimate of σ always tends to its lower bound). With this approach we replace the last term in Eq. (A8) by

$$-0.5\sum\nolimits_b^{-}\log\left\{\exp\left[-u_b w_b\right]+0.01\right\} \qquad \text{(A16)}$$

where $\mathbf{u}=\mathbf{w}^T\mathbf{V}^{-1}$.

To understand how this approach is analogous to the robustification of the multivariate normal likelihood devised by Fournier et al. (1998), we need to consider both the parallels between the forms of the two likelihoods, and the rationale underlying the original robustification. For a single composition, the Fournier et al. (1998) negative log-likelihood, before robustification, may be written as

$$\text{NLL}=0.5B\log(2\pi)+0.5\sum\nolimits_b\log\left[\frac{\xi_b}{N}\right]+\sum\nolimits_b\left[\frac{(O_b-E_b)^2}{2\xi_b/N}\right] \qquad \text{(A17)}$$

where $\xi_b/N=E_b(1-E_b)/N$ is the assumed variance of O_b (I have changed the original notation and regrouped some terms in order to emphasise similarities between this and the logistic-normal likelihood). Because the logistic-normal derives from the multivariate normal, each term in Eq. (A17) has a corresponding term in Eq. (A8): clearly the first terms correspond; also the second and third terms of Eq. (A17) correspond to the third and fourth terms, respectively, of Eq. (A8). The correspondence between the last terms in the two equations becomes more apparent in the special case when there are no correlations in $\mathbf{V}$, in which case the last term in Eq. (A8) could be written as $\sum_b^{-}\left[(Y_b-\mu_b)^2/(2\nu_b)\right]$, where the ν_b are the diagonal terms in $\mathbf{V}^{-1}$ [recall that $w_b=(Y_b-\mu_b)$]. There is no term in Eq. (A17) corresponding to $\sum_b\log(O_b)$ (which is the Jacobian of the transformation Eq. (A4)), but that term needn't concern us because it is effectively a constant (i.e., it includes no likelihood parameters). Thus $\sum_b\log(\xi_b/N)$ is analogous to $\log|\mathbf{V}|$, and w_b is analogous to (O_b-E_b).

Now, the first robustification step of Fournier et al. (1998) was to replace ξ_b by $(\xi_b+0.1/B)$ to avoid this term becoming very small when E_b is very small. This step is not needed with the logistic-normal because $|\mathbf{V}|$ depends only on the weighting parameters, and so does not change when E_b becomes very small. The second step was to replace the third term in Eq. (A17) by $-\sum_b\log\left\{\exp\left[-(O_b-E_b)^2/\left(2\xi_b/N\right)\right]+0.01\right\}$, which ensures that the influence of an observation O_b decreases rapidly as its distance from E_b grows greater than three standard deviations. The analogous modification for the logistic-normal is that given in Eq. (A17) (note that the last term in Eq. (A8) can be written as $0.5\sum_b^{-}u_b w_b$). The final robustification concerned N, which does not appear in the single-composition form of the logistic-normal likelihood, but it does feature – in W_y – in the multi-year form, Eq. (A9). Fournier et al. (1998) replaced N by min(N, 1000) (which implies that all sample sizes greater than 1000 have the same precision). This is a sensible change to make if N represents the number of fish sampled because, as Pennington and Vølstad (1994) showed, effective sample sizes are much smaller than this. However, we may not need this change nowadays because most researchers are already using values of N_y that are much less than the actual sample size. The other point to note is that the N_y in the logistic-normal determine only the relative weighting assigned to each year, whereas the N in the likelihood of Fournier et al. (1998) helps to determine the absolute weighting.

A.5. Parameterizing LN3

LN3 uses the correlation structure of an AR(2) process, which has parameters φ_1, φ_2 and the correlation at lag k, ρ_k, can be calculated recursively by setting $\rho_1=\varphi_1/\varphi_2$ and $\rho_k=\varphi_1\rho_{k-1}+\varphi_2\rho_{k-2}$ (note that $\rho_0=1$ by definition). Since an AR(2) process is stable only if its parameters lie within the triangle defined by $-1\le\varphi_2<1-|\varphi_1|$, it is sensible to reparameterize the LN3 to stop an estimation algorithm considering points outside that triangle. This can be done by replacing the parameters (φ_1, φ_2) by (φ_1, ψ), where $\varphi_2=-1+(2-|\varphi_1|)\,\psi$, and imposing bounds of $(-2, 2)$ and $(0, 1)$ on φ_1 and ψ, respectively. [For LN2 we need only impose bounds $(-1, 1)$ on φ].

A.6. LN3m

Another promising variant of the logistic-normal, discovered after the present paper was in review, is LN3m, which was preferred over LN3 by 52/72 of the data sets of Table 3. This uses the correlation structure of an ARMA(1,1) process (first-order autoregressive, first-order moving average), and has three weighting parameters: σ, φ (for the autoregressive part), and ψ (for the moving average part). The correlation at lag k, ρ_k, can be calculated using $\rho_1=\varphi+\psi/[1+(\varphi+\psi)^2/(1-\varphi^2)]$ and $\rho_k=\varphi^{k-1}\rho_1$ (these equations, and those above for LN3, were derived from the more general "First Method" of calculating autocovariance functions for ARMA processes in Section 3.3 of Brockwell and Davis, 1991). This requires that $-1<\varphi<1$, but ψ is not bounded.

Appendix B. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fishres.2013.12.015.

References

Aitchison, J., 2003. The Statistical Analysis of Compositional Data. The Blackburn Press, Caldwell, New Jersey (this is a revision and updating of the 1986 book of the same authorship and title published by Chapman and Hall).

Aitchison, J., Brown, J.A.C., 1957. The Lognormal Distribution with Special Reference to its Uses in Economics. Cambridge University Press, Cambridge, U.K.

Akaike, A., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19 (6), 716–723.

Brockwell, P.J., Davis, R.A., 1991. Time Series: Theory and Methods, Second ed. Springer, New York.

Bull, B., Francis, R.I.C.C., Dunn, A., McKenzie, A., Gilbert, D.J., Smith, M.H., Bian, R., Fu, D., 2012. CASAL (C++ algorithmic stock assessment laboratory): CASAL User Manual v2.30-2012/03/21. NIWA Technical Report, 135., pp. 280.

Chassot, E., Duplisea, D., Hammill, M., Caskenette, A., Bousquet, N., Lambert, Y., Stenson, G., 2009. Role of predation by harp seals *Pagophilus groenlandicus* in the collapse and non-recovery of northern Gulf of St Lawrence cod *Gadus morhua*. Mar. Ecol. Prog. Ser. 379, 279–297.

Chatfield, C., 2004. The Analysis of Time Series: An Introduction, sixth ed. Chapman & Hall/CRC, Baton Rouge.

Craig, K., 2012. The Beaufort Assessment Model (BAM) with Application to Cobia: Mathematical Description, Implementation Details, and Computer Code. SEDAR28-RW01. SEDAR, North Charleston, SC, pp. 37, Accessed from http://www.sefsc.noaa.gov/sedar/Sedar_Documents.jsp?WorkshopNum=28&FolderType=Review 21 October 2013.

Crone, P.R., Sampson, D.B., 1998. Evaluation of assumed error structure in stock assessment models that use sample estimates of age composition. In: Funk, F., Quinn II, T.J., Heifetz, J., Ianelli, J.N., Powers, J.E., Schweigert J.J., Sullivan, P.J., Zhang, C.I. (Eds.), Fishery Stock Assessment Models: Proceedings of the International Symposium on Fishery Stock Assessment Models for the 21st Century. Anchorage, Alaska, 8-11 October 1998 Alaska Sea Grant College Program Report No. AK-SG-98-01. University of Alaska Fairbanks, pp. 355–370.

Deriso, R.B., Quinn, T.J., Neal, P.R., 1985. Catch-at-age analysis with auxiliary information. Can. J. Fish. Aquat. Sci. 42, 815–824.

Deriso, R.B., Maunder, M.N., Skalski, J.R., 2007. Variance estimation in integrated assessment models and its importance for hypothesis testing. Can. J. Fish. Aquat. Sci. 64 (2), 187–197.

Fournier, D.A., Hampton, J., Sibert, J.R., 1998. MULTIFAN-CL: a length-based, age-structured model for fisheries stock assessment, with application to South Pacific albacore, *Thunnus alalunga*. Can. J. Fish. Aquat. Sci. 55 (9), 2105–2116.

Fournier, D.A., Sibert, J.R., Majkowski, J., Hampton, J., 1990. MULTIFAN a likelihood-based method for estimating growth parameters and age composition from multiple length frequency data sets illustrated using data for southern bluefin tuna (*Thunnus maccoyii*). Can. J. Fish. Aquat. Sci. 47 (2), 301–317.

Francis, R.I.C.C., 2011. Data weighting in statistical fisheries stock assessment models. Can. J. Fish. Aquat. Sci. 68, 1124–1138.

Gazey, W.J., Gallaway, B.J., Cole, J.G., Fournier, D.A., 2008. Age composition, growth, and density-dependent mortality in juvenile red snapper estimated from observer data from the Gulf of Mexico penaeid shrimp fishery. N. Am. J. Fish. Manag. 28, 1828–1842.

Hillary, R., 2011. Bayesian integrated survey-based assessments: an example applied to North Sea herring (*Clupea harengus*) survey data. Can. J. Fish. Aquat. Sci. 68, 1387–1407.

Hirst, D., Aanes, S., Storvik, G., Huseby, R.B., Tvete, I.F., 2004. Estimating catch at age from market sampling data by using a Bayesian hierarchical model. Appl. Stat. 53 (1), 1–14.

Hrafnkelsson, B., Stefánsson, G., 2004. A model for categorical length data from groundfish surveys. Can. J. Fish. Aquat. Sci. 61 (7), 1135–1142.

Kimura, D.K., 1989. Variability, tuning, and simulation for the Doubleday-Deriso catch-at-age model. Can. J. Fish. Aquat. Sci. 46 (6), 941–949.

Kvist, T., Gislason, H., Thyregod, P., 2001. Sources of variation in the age composition of sandeel landings. ICES J. Mar. Sci. 58 (4), 842–851.

Legault, C.M., Restrepo, V.R., 1999. A flexible forward age-structured assessment program. ICCAT Coll. Vol. Sci. Pap. 49 (2), 246–253.

Martell, S., 2011. iSCAM Users Guide Version 1.0, Available from https://sites.google.com/site/iscamproject/. Accessed 21 October 2013.

Martell, S., Schweigert, J., Cleary, J., Haist, V., 2011. Moving towards the sustainable fisheries framework for Pacific herring: data, models, and alternative assumptions. In: Stock Assessment and Management Advice for the British Columbia Pacific Herring Stocks: 2011 Assessment and 2012 Forecasts. Canadian Science Advisory Secretariat Research Document 2011/136, http://www.dfo-mpo.gc.ca/csas-sccs/index-eng.htm/. Abstract accessed 30 November 2013.

Maunder, M.N., 2011. Review and evaluation of likelihood functions for composition data in stock-assessment models: estimating the effective sample size. Fish. Res. 109, 311–319.

Maunder, M.N., Punt, A.E., 2013. A review of integrated analysis in fisheries stock assessment. Fish. Res. 142, 61–74.

McAllister, M.K., Ianelli, J.N., 1997. Bayesian stock assessment using catch-age data and the sampling-importance resampling algorithm. Can. J. Fish. Aquat. Sci. 54 (2), 284–300.

McKenzie, A., 2013. Assessment of hoki (*Macruronus novaezelandiae*) in 2012. New Zealand Fisheries Assessment Report 2013/27., pp. 65.

Methot, R.D., Wetzel, C.R., 2013. Stock synthesis: a biological and statistical framework for fish stock assessment and fishery management. Fish. Res. 142, 86–99.

Miller, T.J., Skalski, J.R., 2006. Integrating design- and model-based inference to estimate length and age composition in North Pacific longline catches. Can. J. Fish. Aquat. Sci. 63 (5), 1092–1114.

Pennington, M., Vølstad, J.H., 1994. Assessing the effect of intra-haul correlation and variable density on estimates of population characteristics from marine surveys. Biometrics 50, 725–732.

Pope, J.G., 1972. An investigation of the accuracy of virtual population analysis using cohort analysis. Res. Bull. Int. Comm. N. W. Atlantic Fish. 9, 65–74.

Punt, A.E., Kennedy, R.B., 1997. Population modelling of Tasmanian rock lobster, *Jasus edwardsii*, resources. Mar. Freshw. Res. 48, 967–980.

R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org. Accessed 21 October.

Rindorf, A., Lewy, P., 2001. Analyses of length and age distributions using continuation-ratio logits. Can. J. Fish. Aquat. Sci. 58, 1141–1152.

Schnute, J.T., Haigh, R., 2007. Compositional analysis of catch curve data, with an application to *Sebastes maliger*. ICES J. Mar. Sci. 64, 218–233.

Schnute, J.T., Richards, L.J., 1995. The influence of error on population estimates from catch-age models. Can. J. Fish. Aquat. Sci. 52, 2063–2077.

Shepherd, J.G., 1999. Extended survivor analysis: an improved method for the analysis of catch-at-age data and abundance indices. ICES J. Mar. Sci. 56 (5), 584–591.

Williams, E.H., Quinn, T.J., 1998. A parametric bootstrap of catch-age compositions using the Dirichlet distribution. In: Funk, F., Quinn II, T.J., Heifetz, J., Ianelli, J.N., Powers, J.E., Schweigert, J.J., Sullivan, P.J., Zhang, C.I. (Eds.), Fishery Stock Assessment Models: Proceedings of the International Symposium on Fishery Stock Assessment Models for the 21st Century. Anchorage, Alaska, 8–11 October 1998, Alaska Sea Grant College Program Report No. AK-SG-98-01. University of Alaska Fairbanks, pp. 371–384.