Classifying multi-frequency fisheries acoustic data using a robust probabilistic classification technique

C. I. H. Anderson and J. K. Horne
School of Aquatic and Fishery Sciences,University of Washington, Box 355020, Seattle, Washington 98195
ciha@u.washington.edu, jhorne@u.washington.edu

J. Boyle
Institute for Systems Biology, 1441 North 34th Street, Seattle, Washington 98103
jboyle@systemsbiology.org

Abstract: A robust probabilistic classification technique, using expectation maximization of finite mixture models, is used to analyze multi-frequency fisheries acoustic data. The number of clusters is chosen using the Bayesian Information Criterion. Probabilities of membership to clusters are used to classify each sample. The utility of the technique is demonstrated using two examples: the Gulf of Alaska representing a low-diversity, well-known system; and the Mid-Atlantic Ridge, a species-rich, relatively unknown system.

PACS numbers: 43.30.Sf, 43.60.Bf, 43.60.Lq [NX]
Date Received: January 23, 2007 **Date Accepted:** March 21, 2007

1. Introduction

Acoustic classifications, discriminations, and identifications of fish and zooplankton species have traditionally integrated prior knowledge, pattern recognition, and direct sampling methods (Horne, 2000). Species identifications have been limited to subjective classification by operators (i.e., scrutinizing; e.g., Dalen *et al.*, 2003) or artificial intelligence (e.g., Haralabous and Georgakarakos, 1996). The use of mean volume backscatter (MVBS) (Kang *et al.*, 2002) and target strength differencing (Gauthier and Horne, 2004) provide objective separations but still assign each target or pixel to a single classification category. The use of multiple frequencies and long-term deployments (e.g., ocean observatories) has increased the need for automated, or semi-automated, data processing techniques, capable of efficient, robust discrimination of ecosystem components and explicit quantification of uncertainty.

Despite the widespread use of probabilistic techniques for prediction and classification in other disciplines [e.g., since the 18th century in weather forecasting (Murphy, 1998)], the certainty of classification has not been included in the analysis of fisheries acoustic data. The increased availability of multi-frequency acoustic data, coupled with ever increasing computing power, facilitates incorporation of probabilistic classification techniques from other fields [e.g., analysis of gene expression data (Boyle, 2005)]. We demonstrate advantages of using unsupervised probabilistic clustering over subjective categorization to classify fish and invertebrates in contrasting ecosystems—low diversity, well-known, and high diversity, relatively unknown.

2. Methodology

2.1 Approach

Probabilistic clustering techniques, such as mixture modeling, differ from partition-based clustering in that each sample is assigned a probability of membership to each cluster rather than an absolute assignment to a single cluster. Partition-based clusters can be described by their centroids (MacQueen, 1967)—the mean position in sample space of all data points assigned to the cluster. In our finite mixture modeling a set of vector models (equivalent to cluster centroids) is

defined, using frequency-specific S_v values, which provide the "best" description of the data. Analytic steps are optimized to allow robust analysis of large data volumes typical of fisheries acoustic data sets (e.g., five frequencies over 250 m depth range for 1 h generate over 23.8 million data values). As samples are assigned probabilities of membership to all clusters, rules must be defined to assign samples to specific clusters.

Applying probabilistic clustering techniques to acoustic data requires three steps: acoustic data processing; generation of clusters and membership probabilities; and analysis of membership probabilities, which includes defining the optimum number of clusters.

2.2 Acoustic data processing

Acoustic data were collected using Simrad EK60 echosounders operating at multiple frequencies between 18 and 200 kHz. Power data were converted to volume backscattering strength (S_v, dB re 1 m^{-1}; cf. Simmonds and MacLennan, 2005), including a range offset of half a pulse length and a transmission loss correction of $20*\log(r)+2\alpha r$ (where r=range from the transducer face and α=frequency-dependent absorption coefficient). Maximum data resolution (i.e., 0.19 m vertical, 1 "ping" horizontal) was retained to maintain the spatial structure of the scattering components. For this analysis, each sample (i.e., image pixel) was assumed to be spatially coincident across frequencies, and no attempt was made to align samples from different transducers. The transducers used to collect the data for both examples were arranged to maximize beam overlap within the physical constraints of transducer placement and beam angles (cf., Korneliussen and Ona, 2002).

2.3 Generation of clusters

Generation of clusters requires the initial selection of the number of clusters, generation of initial vector model values, and then iterative refinement of models. Initial values for vector models were estimated using K-means by median clustering, based on the Euclidean distance measure between S_v values at each frequency, for each sample. Expectation maximization (EM) for finite mixture models (Dempster *et al.*, 1977), where model residuals are based on direct linear distance, was used to refine the models. The expectation step is given by

$$P(x|\mu) = \frac{e^{-\sum_{d\in D}(\mu_d - x_d)^2}}{\sum_{\mu'} e^{-\sum_{d\in D}(\mu'_d - x_d)^2}}, \tag{1}$$

and the maximization step is given by

$$\mu_d = \frac{\sum_{x\in X} x_d P(x|\mu)}{\sum_{x\in X} P(x|\mu)}, \tag{2}$$

where $P(x|\mu)$ is the probability of sample x belonging to model μ, for the set of X samples and D frequencies. The log likelihood (LL) values were approximated using

$$LL = \sum_{x\in X} \log\max(P(x|\mu)). \tag{3}$$

Vector models were iteratively refined until a level of convergence was reached. Convergence was said to occur when the sum of the lowest residuals (i.e., those from the best fitted vector model for each sample) stopped decreasing or a maximum of 15 iterations were completed. Fifteen iterations was chosen as a trade-off between achieving convergence and efficiency in processing large data sets. The posterior probabilities of cluster membership for each sample were then reported. The sum of the probabilities of membership to all clusters is one for each sample.

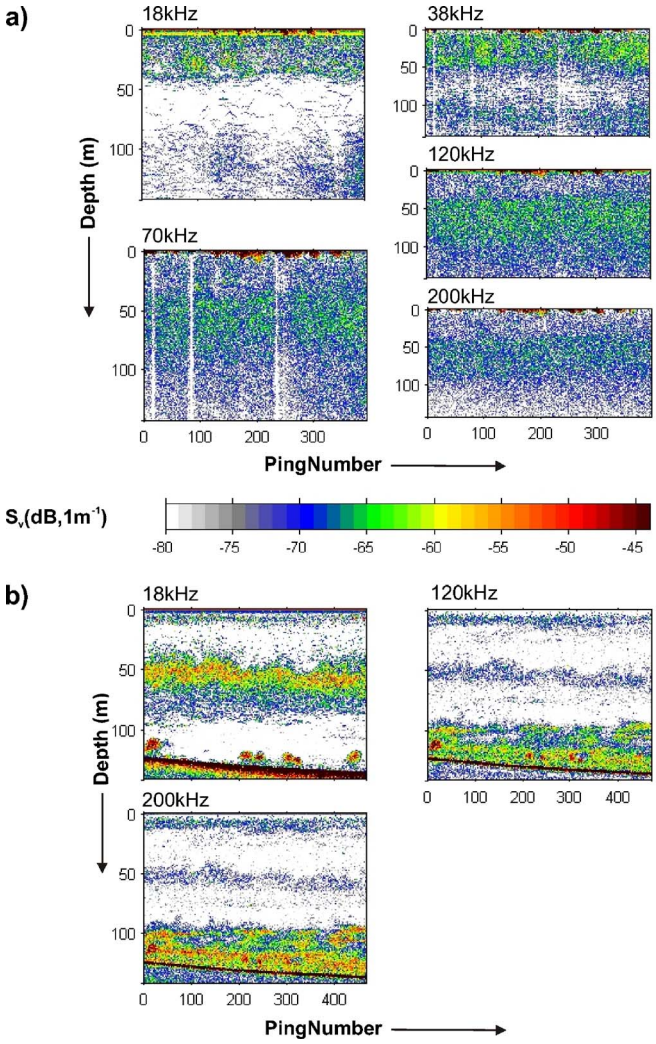

Fig. 1. (Color online) Echograms showing mean volume backscatter (S_v, dB re 1 m^{-1}) data at multiple frequencies from (a) the open ocean above the Mid-Atlantic Ridge (MAR) and (b) the Gulf of Alaska (GoA).

2.4 Analysis of membership probabilities

Probabilities of cluster membership were used to generate synthetic echograms (i.e., probability "maps") to display spatial properties of membership probabilities. Probabilities were also used to assign samples to clusters and as input to cluster metrics. Cluster metrics, including one derived from the Bayesian Information Criterion (BIC) (Schwarz, 1978), were used to determine the optimum number of clusters. Mixture modeling violates the requirements for using the BIC in statistical tests, but versions of the metric are widely acknowledged as useful in assessing the fit of a set of clusters to data (e.g., Fraley and Raftery, 1998). The BIC used here is defined as

$$BIC = -2 * LL + ((M-1) * D) * \log(X * D), \tag{4}$$

where M is the number of models, D is the number of frequencies, and X is the number of samples. BIC values will always be greater than 0, and a higher score represents a poorer description of the data by the vector models. A gradual increase in the score with an increasing

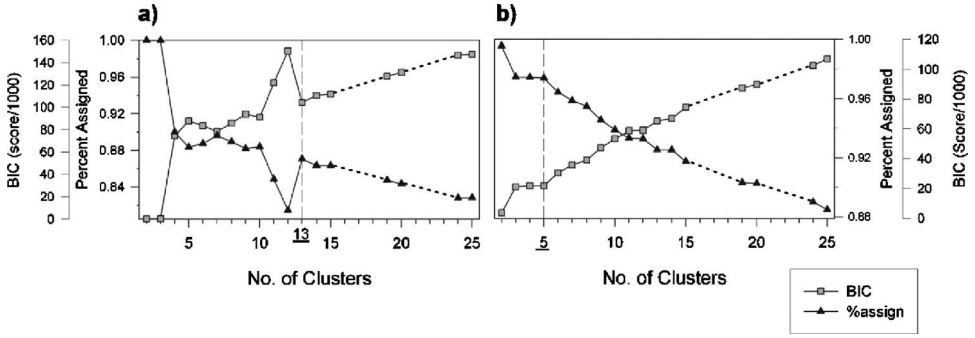

Fig. 2. The Bayesian Information Criterion (BIC) and percentage of clearly assigned samples (%assign) plotted against number of clusters for data from (a) the Mid-Atlantic Ridge (MAR) and (b) the Gulf of Alaska (GoA).

number of clusters is expected as membership probabilities are dissipated among clusters. Large deviations from this upward trend are informative as they represent transitions to another "state." An alternative cluster metric is the percentage of samples clearly assigned to clusters (%assign), where sample x is clearly assigned if $\max(P(x|\mu)) \geqslant 2 * \max - 1(P(x|\mu))$. A higher %assign value represents a better description of the data. The trajectory of the %assign values closely mirrors that of the BIC scores as it is based on the same underlying probabilities.

Selecting the optimum number of clusters is a challenge as the strongest "natural" clusters in acoustic data typically discriminate high S_v values, associated with surface noise and bottom returns where present, from weak returns associated with most biological backscatter. By generating cluster metrics (i.e., BIC or %assign) over a range of numbers of clusters, transitions within the data description can be identified. Interpreting transition points using biological knowledge allows the optimum number of clusters to be identified for each application. An additional benefit of sequentially clustering at an increasing number of clusters is that fidelity of samples to clusters and relationships between clusters can be examined.

3. Examples

3.1 Data

Data for the first example were collected at 18, 38, 70, 120, and 200 kHz during a cruise over the Mid-Atlantic Ridge (MAR) in the North Atlantic [Fig. 1(a)]. This area represents a mid-latitude pelagic ecosystem with a diverse but poorly known epi- and mesopelagic fauna. Echograms are dominated by amorphous horizontal layers and a variety of noise artifacts. The shallowest samples of each ping contain high S_v values due to transducer saturation and intermittent bubbles passing under the transducers. At 38 and 70 kHz vertical white stripes represent "dropped pings" in the data record.

Data were also collected at three frequencies (18, 120, and 200 kHz) in the Gulf of Alaska (GoA) in the northeast Pacific [Fig. 1(b)]. These data are from a continental shelf, high latitude ecosystem with a well-known but limited pelagic fauna. The echograms contain the same transducer saturation feature as seen in the MAR data as well as strong initial returns from the bottom and weaker sub-bottom returns. Biological features include high intensity backscatter from fish schools near the bottom and a series of horizontal layers comprised mainly of forage fish and zooplankton (Stienessen and Wilson, 2002).

3.2 Determining the optimum number of groups

Examination of cluster metrics for the MAR data shows strong transitions at 3 and 13 clusters [Fig. 2(a)]. As expected, 2–3 clusters separate high S_v features (i.e., bubbles and transducer saturation) from low S_v features including biological backscatter (Table 1). This clear discrimination of noise features may be used to remove noise from data sets but does not provide bio-

Table 1. Vector model parameter values (S_v, dB re 1 m^{-1}) defining sets of clusters generated from the Mid-Atlantic Ridge (MAR) and Gulf of Alaska (GoA) data.

Data	Group no.	18 kHz	38 kHz	70 kHz	120 kHz	200 kHz	Comment
MAR- 2	1	−80.81	−74.30	−74.25	−72.12	−75.30	Non-noise features
clusters	2	−37.77	−34.31	−30.28	−28.67	−30.77	Intense noise features
MAR- 3	1	−80.82	−74.32	−74.24	−72.12	−75.29	Non-noise features
clusters	2	−52.49	−48.80	−45.30	−45.14	−49.45	Near-surface bubbles
	3	9.62	12.63	15.78	22.08	24.83	Transducer saturation
MAR- 13	1	−75.34	−72.90	−74.70	−72.66	−87.24	
clusters	2	−91.40	−74.60	−76.20	−73.30	−79.06	
	3	−67.90	−66.20	−70.24	−71.03	−73.76	Includes "fish tracks"
	4	−76.34	−79.87	−72.94	−70.57	−74.14	
	5	−76.54	−72.23	−74.52	−83.90	−76.28	
	6	−61.16	−57.83	−54.80	−53.66	−58.69	Bubble margin+"biota"
	7	−91.29	−87.65	−74.98	−71.01	−73.48	
	8	−71.67	−70.55	−81.48	−72.90	−75.66	
	9	−82.78	−69.58	−70.17	−68.99	−71.61	
	10	−47.33	−43.71	−39.54	−39.93	−43.41	Near-surface bubbles
	11	9.62	12.63	15.78	22.08	24.83	Transducer saturation
	12	−81.96	−81.01	−87.32	−75.56	−78.18	Includes dropped pings
	13	−94.53	−76.06	−69.88	−68.09	−70.49	
GoA- 2	1	−77.44	–	–	−80.95	−79.21	"Low S_v"
clusters	2	−51.70	–	–	−44.83	−42.98	"High S_v"
GoA- 5	1	−27.31	–	–	−38.18	−39.68	1st bottom echo+intense schools
clusters	2	7.33	–	–	21.63	27.32	Transducer saturation
	3	−63.11	–	–	−78.58	−78.16	"Fish"
	4	−85.16	–	–	−90.21	−87.97	"Background"
	5	−86.97	–	–	−68.19	−64.96	"Zooplankton"
GoA- 6	1	−57.63	–	–	−73.47	−72.68	"Large fish"
clusters	2	−22.04	–	–	−32.00	−34.15	1st bottom echo+ intense schools
	3	−87.54	–	–	−68.24	−64.97	"Zooplankton"
	4	−90.99	–	–	−91.53	−88.84	"Background"
	5	−71.70	–	–	−84.76	−84.06	"Small fish"
	6	7.33	–	–	21.64	27.34	Transducer saturation

logically useful resolution of species or species groups. The marked degradation of metric values from 10–12 clusters followed by the strong improvement at 13 clusters suggests a transition in the intrinsic acoustic features described by the clusters. A 13-cluster classification was used to extract the features described beyond the transition point.

Consistent with the MAR data, the "best" description of the GoA data was obtained using two clusters that separated high S_v features, including transducer saturation, bottom echoes, and dense schools, from low S_v values [Fig. 2(b), Table 1]. After a transition in metric values, 3–5 clusters were equivalent in their ability to describe the data, and there was very little structure apparent in metric values above five clusters. A five-cluster classification was chosen for the GoA data to maximize the number of well described biological categories.

3.3 Probabilities of group membership

The 13-cluster classification of the MAR data captures the transducer saturation (cluster 11), intense bubble noise at the surface (cluster 10), and the margins of the bubble features (cluster

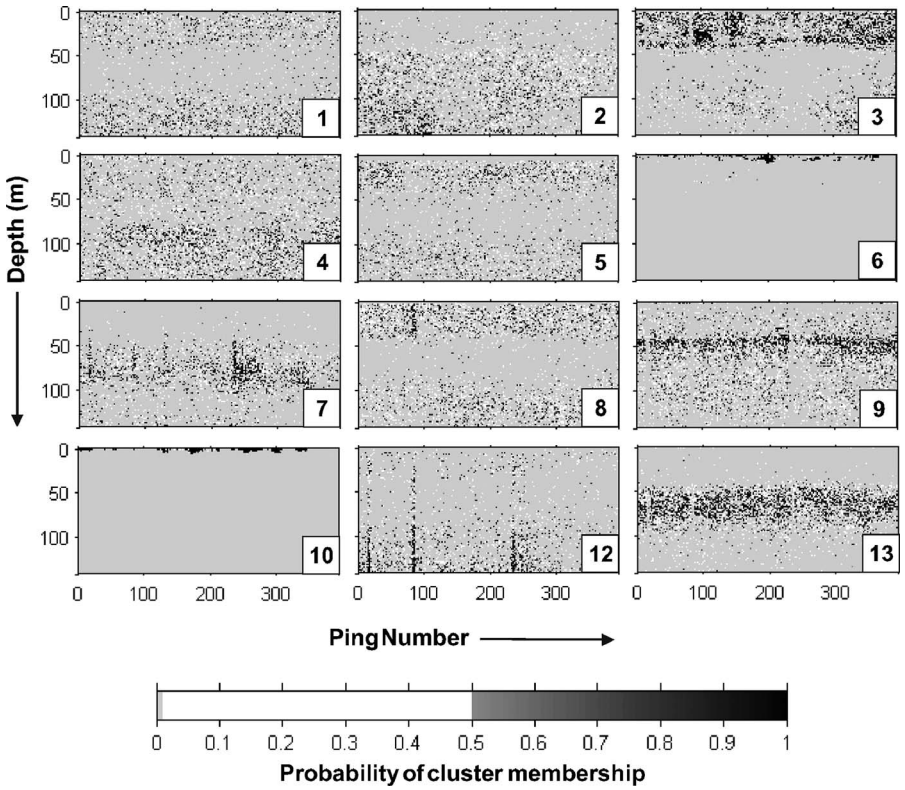

Fig. 3. Probability maps of cluster membership for the 13-cluster classification of Mid-Atlantic Ridge (MAR) data. Notes: Cluster 11 is not shown but is equivalent to cluster 2 shown in Fig. 4. The palest gray tone indicates a membership probability of zero.

6; Table 1, Fig. 3). Dropped pings, along with other samples that have low to moderate S_v values across all frequencies, are found in cluster 12. Similar patterns are evident in clusters 7 and 8. Samples assigned to biological clusters are arranged in horizontal layers, with the three largest clusters (3, 9, and 13) showing contiguous features comprised of high probabilities. Cluster 3 contains individual fish tracks at the same depths as the high probability layer in cluster 13. Vector model parameter values for clusters 3 and 13 show opposite trends in backscatter intensities across frequencies, with cluster 3 containing higher S_v values than cluster 13 at 18 and 38 kHz, and lower values at 120 and 200 kHz, suggesting different types of scattering components (e.g., fish and zooplankton; Table 1). This example demonstrates the ability of objective clustering techniques to extract biological and non-biological features within regions of interest that could not be separated using subjective assignment of contiguous areas within echograms to categories.

GoA probability maps contain a larger percentage of high $\max(P(x|\mu))$ values, which are more spatially contiguous than those in the MAR clusters (compare Fig. 3 and 4). This concentration of high probability values is attributed to both statistical and biological factors. The GoA data are partitioned into 5 rather than 13 clusters resulting in a higher mean $\max(P(x|\mu))$ value (all samples: GoA 5 clusters=0.974, GoA 13 clusters=0.924). However, spatially consistent backscatter intensity patterns, attributed to single species aggregations, also resulted in higher mean $\max(P(x|\mu))$ for clearly assigned samples in the GoA clusters compared to those in the MAR data (clearly assigned samples: GoA 5 clusters=0.989, GoA 13 clusters=0.963, MAR 5 clusters=0.936, MAR 13 clusters=0.927). The presence of spatially

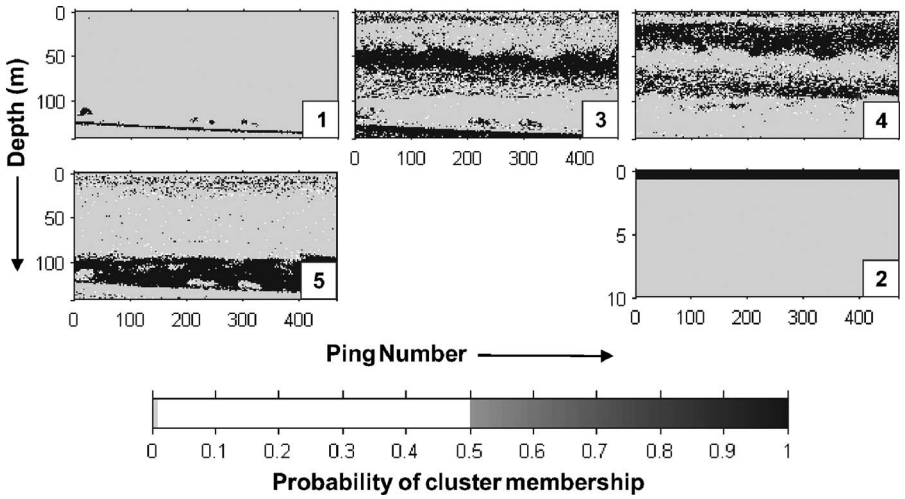

Fig. 4. Probability maps of cluster membership for the five-cluster classification of Gulf of Alaska (GoA) data. Notes: Cluster 2 is drawn on a different depth scale. The palest gray tone indicates a membership probability of zero.

coherent patterns in probability maps from both examples, when no spatial information was used in the clustering process, supports this approach to multi-frequency data classification.

The use of five clusters for the GoA data captures transducer saturation (cluster 2), high intensity backscatter from fish schools and the bottom (cluster 1), and two clusters of horizontally layered biological backscatter with sub-bottom returns (clusters 3 and 5; Fig. 4). The relative magnitudes of vector model S_v values from cluster 3 (higher at 18 kHz and equally low at 120 and 200 kHz) match those from cluster 1 but at lower intensities (Table 1). Frequency-dependent S_v values from cluster 5 are different, lower at 18 kHz and then higher at 120 and 200 kHz. Cluster 4 has relatively low S_v values at all three frequencies, which is interpreted as "empty water" or background backscatter. Results from acoustic and trawl surveys in the GoA (Stienessen and Wilson, 2002) support the biological interpretation of cluster 1 as dense schools of adult walleye pollock, cluster 3 as a mixture of fish at lower densities/sizes (adult and age 0 walleye pollock plus capelin), and cluster 5 as zooplankton (mainly euphausiids). Increasing the number of clusters to six primarily divides cluster 3 into two clusters (1 and 5), with similar relative vector S_v values but at different intensities (Table 1). Inspection of probability maps in conjunction with echograms suggests that this division potentially separates samples containing small from large fish. The use of six clusters to categorize GoA data provides a poorer mathematical description of the data, but may be more appropriate when the objective is to estimate adult walleye pollock biomass independent of other ecosystem components.

4. Future work

The classification of backscatter from contrasting ecosystems demonstrates the potential of probabilistic clustering to analyze multi-frequency fisheries acoustic data. Future work will address: the type of mathematical model used in the EM process, the choice of distance measures to distinguish clusters (including the incorporation of depth and spatial location), and the choice of metrics used to select the optimum number of clusters. Specific issues include the spatial coincidence of samples within transducer beams (Korneliussen and Ona, 2002), and the frequency-dependent loss of signal with depth, which affects choice of EM model. Investigations of spatial and temporal cluster dynamics, including their stability in contrasting ecosystems, will follow refinement of the methodology, with the ultimate goal of automated, robust discrimination of ecosystem components in a wide variety of environments.

Acknowledgments

We thank the MARECO program (www.mareco.no) for providing an opportunity to participate in the MAR cruise and Dr. Alex de Robertis from the NOAA Alaska Fisheries Science Center for providing the GoA data. Two of the authors (C.I.H.A. and J.K.H.) were supported by the Office of Naval Research (NOO14-00-1-0180) and the Alaska Fisheries Science Center (NA17RJ1232-AM01).

References and links

Boyle, J. (**2005**). "Gene-expression omnibus integration and clustering tools in SeqExpress," Bioinformatics **21**, 2550–2551.

Dalen, J., Nedreaas, K., and Pedersen, R. (**2003**). "A comparative acoustic-abundance estimation of pelagic redfish (*Sebastes mentella*) from hull-mounted and deep-towed acoustic systems," ICES J. Mar. Sci. **60**, 472–479.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (**1977**). "Maximum likelihood from incomplete data via EM algorithm," J. R. Stat. Soc. Ser. B (Methodol.) **39**, 1–38.

Fraley, C., and Raftery, A. E. (**1998**). "How many clusters? Which clustering method? Answers via model-based cluster analysis," Comput. J. **41**, 578–588.

Gauthier, S., and Horne, J. K. (**2004**). "Potential acoustic discrimination within boreal fish assemblages," ICES J. Mar. Sci. **61**, 836–845.

Haralabous, J., and Georgakarakos, S. (**1996**). "Artificial neural networks as a tool for species identification of fish schools," ICES J. Mar. Sci. **53**, 173–180.

Horne, J. K. (**2000**). "Acoustic approaches to remote species identification: a review," Oceanogr. **9**, 356–371.

Kang, M., Furusawa, M., and Miyashita, K. (**2002**). "Effective and accurate use of difference in mean volume backscattering strength to identify fish and plankton," ICES J. Mar. Sci. **59**, 794–804.

Korneliussen, R. J., and Ona, E. (**2002**). "An operational system for processing and visualizing multi-frequency acoustic data," ICES J. Mar. Sci. **159**, 293–313.

MacQueen, J. (**1967**). "Some methods for classification and analysis of multivariate observations," in *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics and Probability*, edited by L. LeCam and J. Neyman (University of California Press, Berkeley), pp. 281–297.

Murphy, A. H. (**1998**). "The early history of probability forecasts: Some extensions and clarifications," Weather Forecast. **13**, 5–15.

Schwarz, G. (**1978**). "Estimating dimension of a model," Ann. Stat. **6**, 461–464.

Simmonds, J., and MacLennan, D. (**2005**). *Fisheries acoustics: theory and practice* (Blackwell Science, Oxford), pp. 437.

Stienessen, S., and Wilson, C. (**2002**). "Preliminary results of an interaction study between commercial fishing and walleye pollock (*Theragra chalcogramma*) off East Kodiak, August–September 2002." Cruise number MF2002-009. (Available from RACE Division, Alaska Fish Sci. Cent., NOAA, Natl. Mar. Fish. Serv., 7600 Sand Point Way NE, Seattle, WA 98115.)