Multifrequency species classification of acoustic-trawl survey data using semi-supervised learning with class discovery

M. Woillez,[a] P. H. Ressler, and C. D. Wilson
Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA, 7600 Sand Point Way NE, Seattle, Washington 98115
mathieu.woillez@gmail.com, patrick.ressler@noaa.gov, chris.wilson@noaa.gov

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

Abstract: Acoustic surveys often use multifrequency backscatter to estimate fish and plankton abundance. Direct samples are used to validate species classification of acoustic backscatter, but samples may be sparse or unavailable. A generalized Gaussian mixture model was developed to classify multifrequency acoustic backscatter when not all species classes are known. The classification, based on semi-supervised learning with class discovery, was applied to data collected in the eastern Bering Sea during summers 2004, 2007, and 2008. Walleye pollock, euphausiids, and two other major classes occurring in the upper water column were identified.

PACS numbers: 43.30.Sf, 43.60.Bf, 43.60.Lq [JL]
Date Received: November 17, 2011 **Date Accepted:** December 30, 2011

1. Introduction

Estimating abundance in acoustic-based surveys depends on allocating fish species to measured acoustic backscatter. To help either interpretation by experts or classification by algorithms, trawl samples are collected to provide information (e.g., species, length) about targeted aggregations.[1] However, as most surveys focus on a few commercial species, little direct sampling effort is available to determine species composition of other taxa. Backscatter classification of non-target species may provide added ecological insights.

Species classifications used in fishery acoustics differ in their use of statistical techniques, discriminatory variables, and data spatial resolutions.[2,3] When statistical classification models are built from validation samples and used to predict known species classes, the learning is said to be supervised.[4,5] If no validation samples are used and classes are not specified, then the classification is unsupervised.[6] Classification models can be improved by using both types of information (semi-supervised learning),[7] but detected classes must be present in validation samples. Here a semi-supervised classification procedure is implemented that also allows discovery of unknown classes[8] in multifrequency acoustic data. A generalized Gaussian mixture model (GGMM), adapted to repeated surveys of the same ecosystem, is applied to summer acoustic-trawl survey data in the Eastern Bering Sea (EBS). Walleye pollock (*Theragra chalcogramma*) and euphausiids (*Thysanoessa* spp.) comprise dominant acoustic classes, but other classes with unique frequency responses also occur. Classification results for pollock and euphausiids are compared to results provided by two other classification methods: interpretation by experts[9] and an empirical multifrequency

[a]Also at: School of Aquatic and Fishery Sciences, University of Washington, Box 355020, Seattle, WA 98195.

method.[4] Two other major classes (one known and one unknown) are also identified and described.

2. Multifrequency acoustic and trawl data processing

Acoustic data were collected at 18, 38, 120, and 200 kHz with Simrad EK60 echosounders in the EBS during summers 2004, 2007, and 2008 during NOAA acoustic-trawl surveys.[9] Data acquisition is described in De Robertis *et al.*[4] Background noise at 120 and 200 kHz, and transducer ringing at 18 kHz were estimated.[10] Only data with a signal-to-noise ratio ≥ 10 dB at any frequency, and with $S_v > -80$ dB re 1 m^{-1} at one or more frequencies were used (e.g., Fig. 1). S_v data were averaged into 400 m (horizontal) $\times$ 5 m (vertical) bins. Three pairwise frequency differences relative to 38 kHz (ΔS_v 18-38, ΔS_v 120-38, and ΔS_v 200-38) were used as discriminatory variables for the classification. An existing library of concurrent multifrequency and trawl data characterizing midwater aggregations of single species[4] was used to determine the species-specific pairwise frequency differences. Known acoustic classes included pollock, euphausiids, and a third class of uncertain taxonomic composition showing S_v values highest at 18 kHz (hereafter named high18). Preliminary results[4] suggested that high18 may be jellyfish, but more validation trawls are needed.

3. Classification procedure

Labeled acoustic data, where the acoustic class is known (i.e., library data) and unlabeled data, where class is unknown (i.e., survey data), are learned separately. Then, both models and data are merged together and learned using the GGMM.

3.1 Supervised learning of labeled data

Species-specific pairwise frequency differences were used to determine a classification model, where each known class is assumed to follow a Gaussian distribution.[11] Each class is modeled independently and parameters (mixing coefficients, means, and variance-covariance matrix) are inferred. A single classification model is then built by merging all known class components.

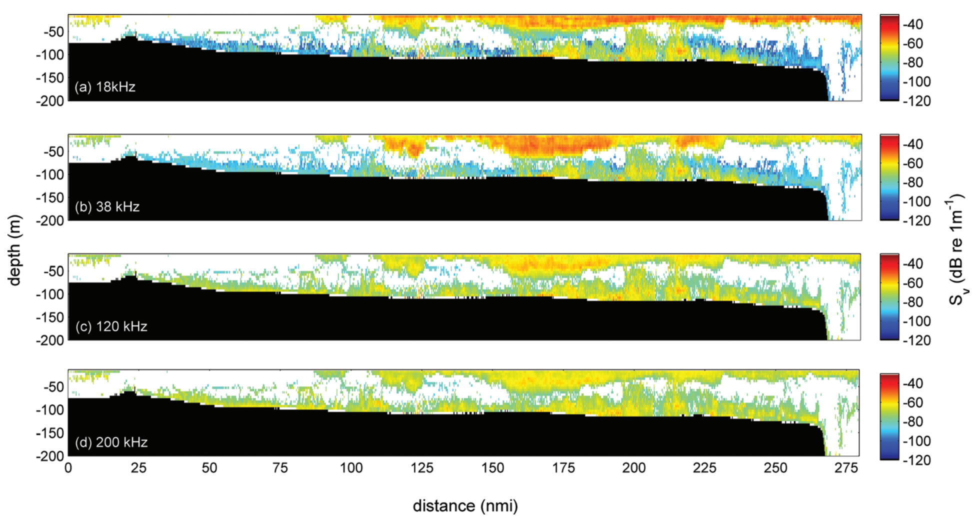

Fig. 1. Volume backscattering strength (S_v in dB re 1 m^{-1}) at 18, 38, 120, and 200 kHz (a, b, c, and d) that met signal-to-noise (SNR ≥ 10 dB at any frequency) and threshold ($S_v > -80$ dB re 1 m^{-1} at one or more frequencies) criteria from representative transect across the EBS shelf, summer 2007.

3.2 Unsupervised learning of unlabeled data

The unlabeled pairwise frequency differences are modeled using a mixture of multivariate Gaussian distributions, where the species class label is an unobserved latent variable that needs to be inferred.[6] To estimate parameters of this statistical classification model, an expectation-maximization (EM) algorithm[12] can be used. It is an iterative method that allows computation of the maximum likelihood estimates of the model parameters. First, a k-means algorithm is used to provide initial estimates of the parameters. Then, the EM iteration alternates between "expectation" (E) and "maximization" (M) steps until convergence is reached. The E-step computes the expectation of the log-likelihood evaluated using current estimates of the parameters, in this case the probability that a data point belongs to one of the Gaussian model components. The M-step computes parameters maximizing the expected log-likelihood found on the E-step. These parameter estimates are then used to determine the distribution of the latent variables (here the class label) in the next E-step. Finally, all unlabeled data are associated with unknown classes.

3.3 Semi-supervised learning based on generalized Gaussian mixture model

All known and unknown class components obtained previously are merged and initialize the semi-supervised GGMM learning.[8] All labeled and unlabeled data are now considered together in the analysis. Thus, some unlabeled data will be classified in known classes. Known class component parameters are then updated. The remaining unlabeled data stay associated with unknown class components, whose parameters are also updated.

The GGMM treats the label presence ($L = l$) or absence ($L = m$) as a random variable. Thus, two types of mixture model components can be defined: (i) predefined components C_{pre}, representing known classes, which generate both labeled and unlabeled data; and (ii) non-predefined components $\bar{C}_{\text{pre}}$, which generate only unlabeled data. These latter components represent outlier regions of known classes or possibly unknown (undiscovered) classes. In addition to mixing coefficients π_k, means μ_k and the variance-covariance matrix Σ_k of Gaussian mixture components, the GGMM introduces three new parameters: (i) the probability of data being labeled when belonging to a generic predefined mixture component M_g, $P(L = l|M_g \in C_{\text{pre}})$ (identical for all predefined components, and equal to the ratio between numbers of labeled and unlabeled data for only data coming from predefined components), (ii) the probability mass function $P(C = c|M_k \in C_{\text{pre}}, L = l)$ associating a probability to each value of the discrete random variable C representing the class value of data when those belong to the kth predefined mixture component and present a label, and (iii) the categorical variable v_k indicating the component type (predefined or non-predefined). The joint log-likelihood of the observed data for the model with N components is then

$$
\begin{aligned}
\log L_N = {} & \sum_{x \in X_l} \log\left(\sum_{k=1}^{N} v_k \pi_k P(x|\mu_k, \Sigma_k) P(L = l|M_g \in C_{\text{pre}}) P(C = c(x)|M_k \in C_{\text{pre}}, L = l)\right) \\
& + \sum_{x \in X_u} \log\left(\sum_{k=1}^{N} v_k \pi_k P(x|\mu_k, \Sigma_k) P(L = m|M_g \in C_{\text{pre}}) + (1 - v_k)\pi_k P(x|\mu_k, \Sigma_k)\right).
\end{aligned}
\tag{1}
$$

Semi-supervised learning uses a generalized EM algorithm to estimate model parameters. It alternates between choosing component types to maximize the log likelihood of the GGMM when all other parameters are fixed, and using a classic EM algorithm to estimate the remaining model parameters when component types are fixed. The choice of the component type is an iterative "one at a time" selection among all possible combinations. That is, a component type is assigned, the model log likelihood evaluated, and the component type yielding the higher likelihood selected. The next component is considered until no further changes are observed. Remaining parameters are estimated using a classic EM, where the expected complete data likelihood is evaluated (E-step), and then model parameters are maximized (M-step). The log likelihood to maximize differs as component types are now determined. Probabilities of component ownership are determined based on whether the data are labeled or unlabeled, and whether a component is predefined or non-predefined. Mixture parameters are then maximized using latter probabilities of ownership.[8]

3.4 Treatment of the year effect and model selection

As dominant acoustic classes are expected to occur repeatedly, a year effect was added to the GGMM framework.[8] Means and variance-covariances of class-components are held constant across years, while mixing coefficients are allowed to vary. The ability of a mixture model to identify classes is quantified using the integrated complete-data likelihood (ICL) criterion, as it includes a penalty to account for classification uncertainty.[13]

4. Results and discussion

4.1 Classification of multifrequency data

Gaussian components were fitted to the labeled data for each of three known classes, and unsupervised learning of unlabeled data was completed using 8 Gaussian components as suggested by the ICL criterion. Then, GGMM classification was used to detect and map 8 classes with different acoustic frequency responses based on four frequency data (Fig. 2).

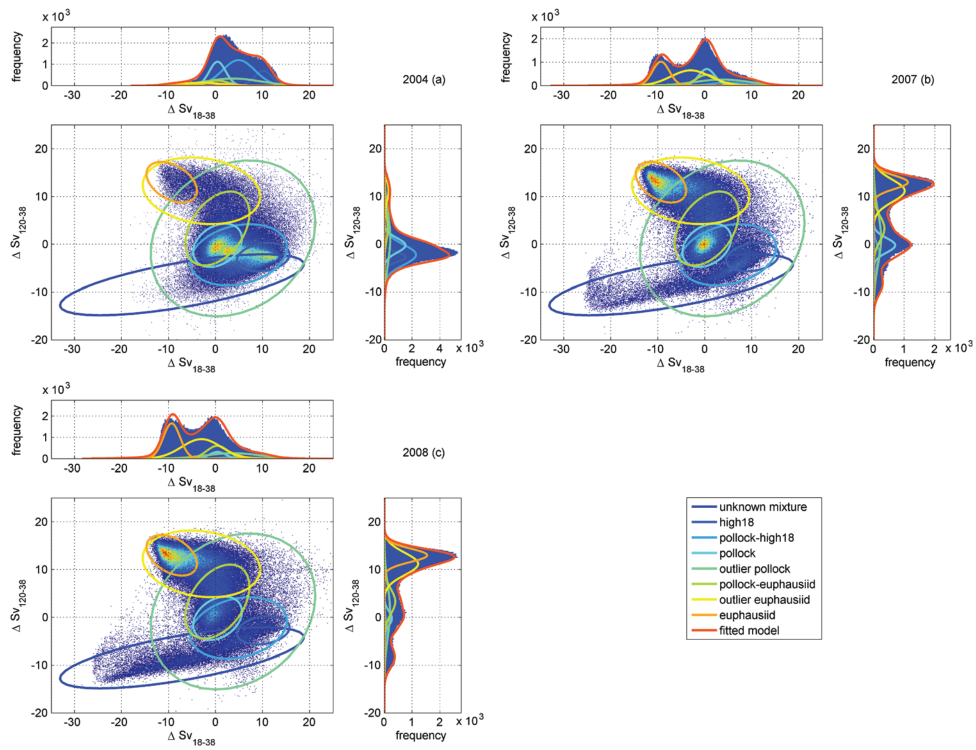

Fig. 2. Example pairwise frequency differences in survey backscatter data for years 2004 (a), 2007 (b), and 2008 (c). 1D and 2D histograms from ΔS_v 18-38 and ΔS_v 120-38 are represented, as well as the eight fitted multivariate Gaussian class components (colors orange to blue). The fitted mixture model, representing the sum of all class components, is represented by a red line in each 1D histogram. The known class components, pollock, euphausiids, and high18, are indicated concurrently with the unknown class components.

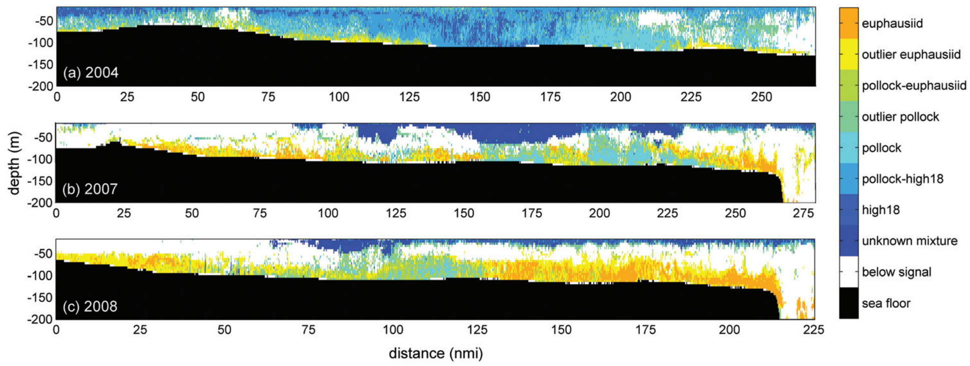

Fig. 3. Corresponding classes obtained by the classification procedure for a representative transect during summer 2004 (a), 2007 (b), and 2008 (c). Known classes are pollock, euphausiid, and high18. Unknown classes correspond to outliers from a known class, mixtures of known classes, or an undiscovered class.

Two classes identified by the GGMM were known classes corresponding to pollock and euphausiids. A third known class, high18, was consistent with the frequency response of jellyfish trawl catches.[4] Two of the five remaining classes were interpreted as outliers distributed around known classes, two as mixtures between two known classes, and one as an undiscovered class, which was different from known classes, outlier distributions, or mixtures. The relative frequency response of the unknown class was similar to that of myctophid trawl catches,[4] but myctophids are not typically found over the EBS shelf. The general shape of the frequency response (peak at 38 kHz) may indicate fish larvae or macroplankton (e.g., siphonophores) containing a small, resonant gas enclosure.[14]

Vertical distribution patterns differed among GGMM scattering classes (Fig. 3). Pollock and euphausiids appeared in the lower two-thirds of the water column during the day (>95% and >97%) in 2007 and 2008, as has been reported elsewhere.[9,15] High18 and the unknown class occurred in the upper third of the water column (>98% and >63%) during these years. However, high18 occurred over most of the water column, overlapping with pollock and euphausiids in 2004.

4.2 Comparison with other classification methods

The GGMM results for two known classes (pollock and euphausiids) and mixtures containing these classes were compared with classification results based on expert classification and/or empirical multifrequency classification for pollock[4] and euphausiids.[15] Mean backscatter attributed to each class for each transect was regressed linearly between methods for each year (Table 1). GGMM results generally agreed with the expert classification of pollock in every year (Table 1) and the empirical multifrequency classification of pollock and euphausiids in two of three years, though in all cases GGMM estimates were lower (slope <1).

For classification challenges associated with mixed aggregations, the three methods produced different results as illustrated in 2004 when the high18 class vertically overlapped with that of pollock and euphausiids on some transects (Fig. 3). The GGMM and expert classification for pollock agreed, but the empirical multifrequency classification pollock results were better correlated with the GGMM pollock-high18 mixture than with GGMM pollock (Table 1). A significant amount of unlabeled data was allocated to a pure class by the empirical multifrequency method, but split into several pure and mixed classes by the GGMM. When combining these two classes, the correlation improved (Table 1). Combining classes also resolved other cases of disagreement (e.g., euphausiids in 2004; Table 1). Indeed, expert and empirical multifrequency classifications aggregated around target classes, while the GGMM classification tended to subdivide in more classes than only the target ones.

Table 1. Class estimates of alternative classification methods for pollock and euphausiid were compared to corresponding GGMM class estimates by linear regression. Transect mean backscatter was used as the comparative variable. From all possible comparisons ($n = 51$), only significant ($\alpha < 0.001$) models with a high r^2 and a slope greater than 0.5 are shown. k: for known class component.

Alternative methods class estimates	GGMM class estimates	year	intercept	slope	r^2
Expert classification for pollock	pollock (k)	2004	35.80	0.733	0.850
		2007	−1.54	0.685	0.888
		2008	−12.02	0.513	0.861
	pollock (k) + outlier pollock	2004	53.934	0.794	0.862
		2007	3.381	0.746	0.906
	pollock (k) + pollock-euphausiid	2008	−17.054	0.770	0.909
Empirical multifrequency Classification for pollock (Ref. 4)	pollock-high18	2004	−169.33	1.013	0.741
	pollock (k)	2007	−21.70	0.737	0.951
		2008	−16.66	0.552	0.903
	pollock-high18 + pollock (k)	2004	121.029	1.039	0.891
	pollock (k) + outlier pollock	2007	−17.725	0.799	0.960
	pollock (k) + pollock-euphausiid	2008	−22.639	0.818	0.930
Empirical multifrequency Classification for euphausiids (Ref. 13)	outlier euphausiid	2004	−12.53	1.099	0.887
	euphausiid (k)	2007	−30.88	0.821	0.956
		2008	−14.03	0.699	0.937
	outlier euphausiid + euphausiid (k)	2004	−10.54	1.180	0.949
		2007	−11.536	0.903	0.970
		2008	12.883	0.770	0.939

5. Conclusion

Semi-supervised learning using a GGMM effectively classified multifrequency acoustic backscatter from known and unknown species classes. The GGMM results for known scattering classes, pollock and euphausiids, compared well with classification results from established methods, although differences occurred when backscattering types were mixed. The GGMM was able to create mixtures of classes, whereas expert scrutiny and empirical multifrequency classification assume pure classes. The main advantages of the GGMM method are that (i) it can discover unknown classes while accounting for the presence of known classes and (ii) it can include mixtures between identified classes. To improve the GGMM method, future work should include labeled data based on mixed aggregations, other discriminatory variables (e.g., morphological descriptors, depth), and direct sampling of unknown scattering classes.

Acknowledgments

Support was provided by Alaska Fishery Science Center, NMFS, NOAA, and the Bering Sea Integrated Ecosystem Research Program (publication #34), North Pacific Research Board (publication #327). We thank Alex De Robertis for supplying labeled data and reviewing the manuscript, Jim Ianelli and two anonymous referees for manuscript review, and Ainhoa Lezama-Ochoa for helpful discussions. Findings and conclusions are those of the authors and do not necessarily represent views of NMFS, NOAA. Reference to trade names does not imply endorsement.

References and links

[1]J. Simmonds and D. MacLennan, *Fisheries Acoustics: Theory and Practice* (Blackwell Science, Oxford, 2005).

[2]J. K. Horne, "Acoustic approaches to remote species identification: A review," Fish. Oceanogr. **9**, 356–371 (2000).

[3]P. G. Fernandes, "Classification trees for species identification of fish-school echotraces," ICES J. Mar. Sci. **66**, 1073–1080 (2009).
[4]A. De Robertis, D. R. McKelvey, and P. H. Ressler, "Development and application of an empirical multifrequency method for backscatter classification," Can. J. Fish. Aquat. Sci. **67**, 1459–1474 (2010).
[5]R. J. Korneliussen, Y. Heggelund, I. K. Eliassen, and G. O. Johansen, "Acoustic species identification of schooling fish," ICES J. Mar. Sci. **66**, 1111–1118 (2009).
[6]C. I. H. Anderson, J. K. Horne, and J. Boyle, "Classifying multifrequency fisheries acoustic data using a robust probabilistic classification technique," J. Acoust. Soc. Am. **121**, 230–237 (2007).
[7]K. Nigam, A. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using EM," Mach. Learn. **39**, 103–134 (2000).
[8]D. J. Miller and J. Browning, "A mixture model and EM-based algorithm for class discovery, robust classification, and outlier rejection in mixed labeled/unlabeled data sets," IEEE T. Pattern Anal. **25**, 1468–1483 (2003).
[9]T. Honkalehto, W. Patton, S. De Blois, and N. J. Williamson, "Echo integration trawl survey results for walleye pollock (Theragra chalcogramma) on the Bering Sea shelf and slope during summer 2000," NOAA Technical Memorandum No. NMFS AFSC-126 (2002).
[10]A. De Robertis and I. Higginbottom,"A post-processing technique to estimate the signal-to-noise ratio and remove echosounder background noise," ICES J. Mar. Sci. **64**, 1282–1291 (2007).
[11]G. J. McLachlan, *Discriminant Analysis and Statistical Pattern Recognition* (Wiley, New York, 1992).
[12]A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via EM algorithm," J. R. Stat. Soc. Ser. B (Methodol.) **39**, 1–38 (1977).
[13]C. Biernacki, G. Celeux, and G. Govaert, "Assessing a mixture model for clustering with the integrated completed likelihood," IEEE T. Pattern Anal. **22**, 719–725 (2000).
[14]M. V. Trevorrow, D. L. Mackas, M. C. Benfield, "Comparison of multifrequency acoustic and in situ measurements of zooplankton abundances in Knight Inlet, British Columbia," J. Acoust. Soc. Am. **117**, 3574–3588 (2005).
[15]P. H. Ressler, A. De Robertis, J. D. Warren, J. N. Smith, and S. Kotwicki, "Developing an acoustic index of euphausiid abundance to understand trophic interactions in the Bering Sea ecosystem," Deep Sea Res. II (in press).