Steven T. Hoekman. 2021. Multi-observer methods for estimating uncertain species identification. *Ecosphere*.

Metadata S1: Example R code for estimating uncertain species identification using multi-observer methods

Author

Steven T. Hoekman
Wild Ginger Consulting
P. O. Box 182
Langley, WA, USA 98260
steven.hoekman@protonmail.com

File list (found within DataS1-Example_R_Code)

R code files

`example_r_code.R`	R code for example statistical analyses
`example_data.RData`	R workspace with data, functions and output
`likelihood_equations_a2b2g25.RData`	Pre-computed values for computations

Dependencies

R packages (available at <https://cran.r-project.org/web/packages/>)

`plyr` (required)	Programming tools, data manipulation
`dplyr` (required)	Data frame manipulation
`mrds` (optional)	Delta method for arbitrary functions

Files description

`example_r_code.R`

An R file containing packages, functions, and computer code for statistical analyses estimating uncertain species identification using multi-observation method (MOM) and single-observation method (SOM) models. Functions ‘example.1.f’ to ‘example.5.f’ conduct analyses of example survey data. Also included are required supplemental functions. These functions were also used in simulation analyses (DataS3-R_Code_for_Simulations), with some simplified for use here. Detailed comments in code describe their inputs, outputs, and functioning.

`example_data.RData`

An R workspace containing example data, example statistical output, R objects, and R functions for conducting example statistical analyses.

R data frames ‘data_1’ to ‘data_5’ contain example survey data. Columns names beginning with ‘y_’ denote counts of individuals classified by each observer (with ‘p’ specifying primary observers and ‘s1’, ‘s2’, . . . specifying different secondary observers) to each observation state (with an integer after an underscore specifying observation state *a*). Column ‘group_size’ is the count of individuals in the observed group; example functions require data are sorted by ascending ‘group_size’. Column names beginning with ‘covariate’ specify group-level, standard normal covariates predicting either true species probabilities (‘_psi’) or classification probabilities (‘_theta’). Each row below the column names contains a unique observation history (a unique set of observed groups and covariate values), with the column ‘count’ enumerating the count of each unique observation history.

R lists ‘model_output_1’ to ‘model_output_5’ provide statistical output for each example.

The workspace also includes R functions in the `example_r_code.R` file and the R list ‘likelihood.equations’ provided in the `likelihood_equations_a2b2g25.RData` file.

`likelihood_equations_a2b2g25.RData`

An R list object containing the list object ‘likelihood.equations’ required to provide pre-computed values for likelihood equations calculating probabilities for observed groups. See MetadataS3.pdf in DataS3-R_Code_for_Simulations for further details.

Examples description

Five examples demonstrate statistical analyses using the R statistical computing environment (R Development Core Team 2020) for a diverse set of models using multi-observer methods (Table 1). Example code favors clarity over flexibility and computational efficiency. Analyses for each example require supplemental functions and files, which are listed in the comments with each example and provided in the R workspace `example_data.RData`. A consistent naming convention is used for all estimated parameters (Table 2).

These relatively simple example models should optimize reliably. However, I included example code using box constraints and penalty functions, which can help to avoid unrealistic or inadmissible parameter values when optimizing more complex models.

Models use the R ‘optim’ function to minimize the –log(likelihood) using the L-BFGS-B optimization method of Byrd et al. (1995), which allows box constraints giving lower and upper bounds to values of individual parameters. For all models, I included appropriate lower box constraints, which aided optimization. Upper box constraints often slowed optimization, and I don’t recommend these unless necessary to avoid optimization problems or unrealistic parameter values (such as can occur with multimodal optimization). Because these issues are more prevalent with SOM models, example 2 includes upper box constraints.

A flexible alternative approach is to constrain optimization using a penalty function. These models assume that probabilities are bounded between 0 and 1 and that groups of complementary probabilities (e.g., true species probabilities) sum to 1. Logit transformation of probability parameters satisfies the former assumption, but not the latter. In example 1, if true species probability parameters for species 1 and 2 sum to greater than 1, probability for species 3 (defined as 1 – the summed probabilities for species 1 and 2) is negative. A penalty function defines a penalty term taking value 0 if summed probabilities for species 1 and 2 are ≤ 1, but otherwise takes value >0 that scales with deviation from 1. Adding a penalty term to the −log(likelihood) avoids optimization that violates the constraint. In example 4, a penalty term enforces the constraint that heterogeneous group probability (π_{12}) does not take inadmissible values (i.e., when heterogeneous groups for a species would exceed its total number of groups).

Examples 2 and 4 include groups of size ≥ 1. Because likelihood calculations become increasingly complex and computationally intensive for larger groups, these examples require the list ‘likelihood.equations’ provided in `likelihood_equations_a2b2g25.RData`. This list increases computational efficiency by supplying pre-computed values for likelihood equations. See MetadataS3.pdf in DataS3-R_Code_for_Simulations for further details.

Lists ‘model_output_1’ to ‘model_output_5’ provide sample statistical output for each example. Slight differences in output may result from differing initial parameter values or box constraints, etc., but difference should not be substantive. Output is a list with elements including: parameter estimates (‘par’), the −log(likelihood) + any penalty term(s) (‘value’), and the Hessian matrix for estimation of the variance-covariance matrix (‘hessian’).

Workflow for executing example analyses

To execute statistical analyses for these examples:

1) Unzip the .R and .RData format files in DataS1 to the R working directory.
2) In an R editor or integrated R development environment, open the file `example_r_code.R`
3) Install and load required (and optional) R packages.
4) Load the R workspace `example_data.RData`

   The workspace includes data, functions, and output for example analyses.
5) Navigate to the desired example code section and generate a vector of initial parameter values (‘parameters_ini_1’ to ‘parameters_ini_5’).

   Initial values for probability parameters (ψ, θ, π) are entered as probabilities and subsequently logit-transformed; values for group size parameters are average group size.
6) Generate vector(s) of lower (‘constraints_low_1’ to ‘constraints_low_5’) and upper (optional) box constraints for parameter values.

   Only example 2 includes upper box constraints (‘constraints_up_2’).
7) Execute the appropriate ‘optim’ R function code to produce a list (‘model_1’ to ‘model_5’) with statistical output.

   More complex examples may require several minutes to optimize.
8) If desired, apply the ‘plogis’, ‘DeltaMethod’, and ‘solve’ functions to estimate variances and standard errors for estimated parameters and transformed parameter estimates (e.g., back-transformed logit parameter estimates).

   See code provided in example 1.

Literature Cited

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. 1995. A limited memory algorithm for bound constrained optimization. SIAM Journal on Scientific Computing. 16:1190–1208.

R Development Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. www.r-project.org.

Table 1. MOM and SOM models used in examples. Model notation and definitions of parameters and terms follows descriptions in the companion article and Appendix S2. Models with covariates include a continuous, group-level, standard normal covariate predicting either true species probabilities or classification probabilities *via* a multinomial logit regression including an intercept coefficient and a slope coefficient. Identical secondary observers are assumed to have identical classification probabilities, but classification probabilities differ among distinct secondary observers.

Example	Model	Observation states A	True species States B	Groups	Covariate	Primary observer	Secondary observers
1	$\mathbf{M}^{3\|3}$	3	3	Size 1	-	1	3 identical
2	$\mathbf{S}_{g\geq 1}^{2\|2}$	2	2	Homogeneous	-	-	4 identical
3	$\mathbf{M}_{\psi}^{3\|2}$	3	2	Size 1	True species probability ψ	1	4 identical
4	$\mathbf{M}_{\mathrm{het}(g\geq 1),\theta}^{2\|2}$	2	2	Heterogeneous	Classification probability θ	1	4 identical
5	$\mathbf{M}_{\{\mathrm{s}\}}^{2\|2}$	2	2	Size 1	-	1	3 distinct

Table 2. Naming conventions for parameters in example models. Parameters have a base name (bold) specifying the type of parameter (column 1). Suffix(es) are appended to denote observation state, true species state(s), and observer identity. For multinomial logistic regression coefficients predicting true species probabilities or classification probabilities, a prefix denotes intercept (b0) versus slope (b1) coefficients. Suffixes and prefixes are separated from the base name and each other by an underscore. Parameters are described in the companion article.

Parameter type (notation)	Name	Prefix	Suffix(es)
Heterogeneous groups (π)	**pi**_[*ab*]		Two integers denote true species states *b*
Classification probabilities (θ)	[beta]_**theta**_[observer]_[*ab*]	Text ‘b0’ or ‘b1’ denotes intercept or slope regression coefficients for multinomial logistic regression predicting classification probabilities	1) [observer] Text denotes primary ‘p’ or secondary observers[†] ‘s1’, ‘s2’, . . . 2) [*ab*] Two integers denote classification state *a* and true species state *b*
True species probabilities (ψ)	[beta]_**psi**_[*b*]	Text ‘b0’ or ‘b1’ denotes intercept or slope regression coefficients for multinomial logistic regression predicting true species probabilities	Integer denotes true species state *b*
Mean group size ($\bar{g}$)	**g**_[*b*]		Integer denotes true species state *b*

[†] If only one secondary observer is specified, secondary observers are assumed to have identical classification probabilities.