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 folder 'DataS1-Example_R_Code')

R 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 functions and computer code for statistical analyses estimating uncertain species identification using multi-observation method (MOM) and single-observation method (SOM) models. Functions include '`example.1.f`' to '`example.5.f`' for conducting analyses of five example data sets and required supplemental functions. Most are simplified versions of those used in simulation analyses (found in folder 'DataS3-R_Code_for_Simulations'). Detailed comments in code describe their inputs, outputs, and functioning.

`example_data.RData`

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

R data frames '`data_1`' to '`data_5`' contain five examples of survey data. Columns with names beginning with 'y_' contain counts of individuals classified by each observer (with 'p' specifying the primary observer and 's1', 's2', . . . specifying 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. **Data are required to be sorted by ascending 'group_size'**. Column names beginning with 'covariate' specify group-level, standard normal covariates predicting either true species probabilities (suffix '_psi') or classification probabilities ('_theta'). Rows contain distinct observation histories (a distinct set of observed groups and covariate values), with column 'count' enumerating the count of each distinct observation history.

R lists '`model_output_1`' to '`model_output_5`' contain statistical output for each example.

The workspace also includes R functions provided 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 workspace containing the list object '`likelihood.equations`' required to provide pre-computed values for likelihood computations for models with group size ≥1. See MetadataS3.pdf in folder 'DataS3-R_Code_for_Simulations' for further details.

Examples description

Five examples demonstrate statistical analyses using the R statistical computing environment version 3.6 (R Development Core Team 2020) for a diverse set of multi-observer method models (Table 1). Example code favors clarity over flexibility and computational efficiency. Supplemental functions and files required for each analysis are listed in the comments before

each example and are provided in the R workspace `example_data.RData`. Table 2 provides parameter-naming conventions.

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 negative log(likelihood) using the L-BFGS-B method of Byrd et al. (1995), which allows box constraints placing lower and upper bounds on values of individual parameters. For all models, I included appropriate lower box constraints, which aided optimization. Upper box constraints often slowed optimization. I don't recommend these unless necessary to avoid optimization problems or unrealistic parameter estimates (such as can occur with multi-modal optimization). Because these issues were more prevalent with SOM models, example 2 included 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. With ≥2 complementary probabilities, logit transformation of parameters satisfies the former assumption, but not the latter. In example 1, if true species probabilities for species 1 and 2 sum to >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 taking value >0 that scales with deviation from 1. Adding a sufficiently large penalty term to the negative log(likelihood) discourages optimization violating the constraint. In example 4, a penalty term enforces a 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 statistical output for each example. Slight differences in your 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 negative 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) Place the .R and .RData format files within DataS1 into the R working directory.
2) In an R editor or integrated R development environment (e.g., *RStudio* by RStudio, Inc., Boston, Massachusetts, USA, <https://rstudio.com>), 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 example data, functions, and output from analyses.
5) Navigate to the desired example code section and generate initial parameter values ('`parameters_ini_1`' to '`parameters_ini_5`').
    - Enter initial values for probability parameters (ψ, θ, π) as probabilities (they are subsequently logit-transformed); enter initial values for group size parameters (g) as mean group size.
6) Generate 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' function 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 for transformed parameter estimates (e.g., parameters back-transformed from logit-scale to probability-scale).
    - 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, Appendix S1, 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 with log-linear predictors including an intercept coefficient and a slope coefficient. Identical secondary observers have the same 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	$\mathbf{S}^{2,2}_{g\geq 1}$	2	2	Homogeneous	-	-	4 identical
3	$\mathbf{M}^{3,2}_{\psi}$	3	2	Size 1	True species probability ψ	1	4 identical
4	$\mathbf{M}^{2,2}_{\theta,\mathrm{het}(g\geq 1)}$	2	2	Heterogeneous	Classification probability θ	1	4 identical
5	$\mathbf{M}^{2,2}_{\{s\}}$	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, Appendix S1, and Appendix S2.

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 parameters for one secondary observer are specified, secondary observers are assumed identical (i.e., to have the same classification probabilities).