Stulletal_mcp Quick Reference Sheet

Contents

1 Introduction

This document serves as a quick reference sheet for the GitHub repository **stulletal_mcp**. Section 1 provides a brief introduction and overview of the authors' intentions and goals for the mixed cumulative probit (MCP) model, as well as guidelines for researchers to use the MCP and apply the modeling technique to new data and research questions. Section 2 explains different applications of the repository using new data and detailed instructions on script modifications. For more information on the MCP, its inception, and practical application for subadult age estimation, readers can refer to the original publication, "Mixed Cumulative Probit: A novel algorithm inspired by methodological and practical shortcomings of age estimation in biological and forensic anthropology" (Stull et al., *In review*).

1.1 Mixed Cumulative Probit Overview

The impetus for creating the mixed cumulative probit (MCP) was to model continuous processes (*e.g.*, growth) using a combination ("mixture") of one or several continuous, categorical, and/or ordinal observations based on Bayes' Theorem of probability. Both multivariate (using multiple response variables) and univariate (a single response variable) models can be generated using the back-end parameter optimization (found in scripts *solvey_US_multivariate.R* and *solvey_US_univariate.R*, respectively) and model selection functions (found in scripts *make_multivariate_crossval_results.R* and *make_univariate_crossval_results.R*, respectively) included in the pipeline outlined by the **stulletal_mcp** repository. Training a complex, multivariate model necessitates an iterative process that allows for multiple instances of testing and updating variable parameters. The present version of the MCP uses Hooke-Jeeves algorithm to perform parameter optimization and maximum likelihood estimation for model selection.

1.2 Mixed Cumulative Probit Applications

If some parts of the overview section and/or the Supplemental Information document for the corresponding article presenting the MCP and its application for age estimation in biological anthropology have gone over your head, there is no need to panic! Understanding the complexity behind the MCP is in no way necessary to apply it. To this end, the authors have made the MCP available for use via three different interfaces, each with its own advantages and limitations, levels of accessibility, and application capabilities.

1. **KidStats** Pre-optimized models for subadult skeletal age estimation will be made available through the newest version of KidStats (v.2.0.0), a freely available Shiny® (Chang et al., 2021) application accessible at `https://kyra-stull.shinyapps.io/kidstats/` or as an R-package through GitHub (`https://github.com/ElaineYChu/kidstats`) using RStudio® (RStudio Team, 2021). KidStats is for researchers who wish to run the MCP on a single individual or

sample through already-constructed MCP models. Advantages to using KidStats include time saved by not having to optimize models and the ability to decide which variables to include in the analysis (either univariate or multivariate). Assumptions for using this interface are that the individual and/or sample being run through KidStats match the sample demographics used to optimize the model (*i.e.*, a sample of 1325 US individuals, assumed to be between the ages of 0 and 20 years). Researchers would most likely use KidStats for subadult skeletal age estimation using the response variables (one, all, or a subset) and the same protocols for measuring and scoring them that were originally included to build the model.

2. **stulletal_mcp** This GitHub repository (https://github.com/ElaineYChu/stulletal_mcp) hosts code that is set up as a pipeline of six scripts that can be sourced in R for a step-by-step replication of the steps used to generate results for the main article. Each script contains detailed descriptions about the mixed cumulative probit pipeline and the functions housed within. This interface provides a major advantage to researchers who wish to optimize the mixed cumulative probit model for their own data but may not have the R proficiency to recall their own pipeline using the functions themselves. The repository is set up such that a researcher interested in using the mixed cumulative probit to build subadult skeletal age estimation models based on their own samples may do so with minimal edits to the pre-existing scripts (**Section 2**). Any user editing the code provided by the GitHub repository should know they are relying on the pipeline already defined by the existing scripts. Therefore, language and application of these scripts correspond to the specific research question of subadult age estimation using skeletal and dental age indicators.

3. **R Package *yada*** The R package *yada* is also housed on GitHub and is publicly available for installing via R at https://github.com/MichaelHoltonPrice/yada. All functions for performing model optimization, selection, evaluation, and visualization are included in *yada*. The functions in this package allow the user to apply the mixed cumulative probit to their own appropriate research questions. For best results, this approach requires some proficiency in using R, and an above-average understanding of the processes involved in optimizing, selecting, and applying the MCP model. Examples of other applications using the mixed cumulative probit include modelling any continuous phenomenon using multiple (ordinal and/or continuous) observations in a Bayesian framework.

The three interfaces available to apply the MCP were created to maximize accessibility of this new modeling technique for researchers with various levels of statistical and computational experience. Those who wish to treat the functionalities of the MCP as a sort of "black-box," or are satisfied with an overarching understanding of what the models are and their use, may want to select either the first or second interfaces provided. Akin to FORDISC® (Jantz & Ousley, 2005) commonly used by forensic anthropologists, **KidStats** may be used to estimate the age of a single subadult individual using one, multiple, or all common skeletal and dental growth indicators included in the Shiny® application. Also similar to FORDISC®, a user may apply the MCP's process of parameter optimization and model selection on their own data for modeling continuous processes using the pipeline established in the **stulletal_mcp** repository. Those who have taken the time to review each available function and feel comfortable with the processes made available in the **R Package *yada*** are invited to take advantage of the flexibility to apply the functions to their own data and research questions in ways that the authors have not yet considered. By providing multiple avenues through which the MCP can be applied, the authors look forward to creative and innovative research in the future.

2 Using the Repository for New Data

The **stulletal_mcp** GitHub repository serves two distinct purposes:

1. Make all code and the corresponding article (Stull *et al.*, 2021) freely accessible
2. Provide code that can be manipulated and adapted by researchers for the same purpose using their own (new) data

---

Two methods of manipulating the **stulletap_mcp** repository for one's own research are provided below:

> **Changing file names**
> The simplest method for using the repository scripts is to move the new dataset into the *data* folder and rename the data file as "*SVAD_US.csv*" and provide the variable information file and name it as "*US_var_info.csv*". Make sure that the general format for both new files match those of the original files. In particular, the variable information file must contain the required four (4) or eight (8) columns that are described in the stulletlal_mcp GitHub README file.
>
> **Editing Scripts**
> Another method for using the repository is to edit the individual R scripts. This method allows for greater flexibility on the user's end to manipulate the code as needed for new analyses. Within each script, file locations and data file names can be changed to match the user's new data. In general, the "*analysis_name*" ("US" in the original files and scripts) may be changed for all script names, as well as the folder name in which all scripts are housed. If using this method, be sure that "data" and "results" folders exist as subfolders. The tables on the next two pages describe the lines and changes required for this approach by script name.

In addition to changing individual scripts, the pipeline may also be slightly altered, depending on individual needs:

- Variable names and the number of response variables can be altered
- If the researcher is only interested in creating univariate models, the script *solvey_US_multivariate.R* can be skipped.
- Alternatively, if the researcher is only interested in creating multivariate models, *solvey_US_multivariate.R* can generate multivariate models without univariate models being created first, thus allowing the researcher to skip the *univariate* script.

Script Name	Line Number(s)	Change
write_US_problems.R	80-83	Change "stulletal_mcp" to new folder name and set as your working directory
	88	Change "US_var_info.csv" to new variable information file name
	89	Change "SVAD_US.csv" to new data file name
	100	Change "US" to desired analysis name
	108	You may change the number of test and training folds (4) to desired number ("K=") and change the seed number ("seed=")
solvey_US_univariate.R	57	Currently, this analysis uses all available cores on the local system to perform calculations. You may change this by replacing "detectCores()" with "cl=makeCluster(*number of cores here*)"
	62	Change "US" to desired analysis name
	73	You may change the value of "base_seed"
make_univariate_crossval_results.R	30	Change "US" to desired analysis name
	37-39	You may change variables "cand_tol", "scale_exp_min", and "beta2_max" to values that are appropriate for your own analysis
solvey_US_multivariate.R	32	Currently, this analysis uses all available cores on the local system to perform calculations. You may change this by replacing "detectCores()" with "cl=makeCluster(*number of cores here*)"
	37	Change "US" to desired analysis name
multivariate_fit_wrapper Option 1 - Correlation Groups	41-112	Currently, the pipeline assumes four correlation groups to speed-up optimization, but you may alter lines 84-93 to establish different response variable groupings
multivariate_fit_wrapper Option 2 - No Correlation Groups	41-112	If you have no assumptions on response variable groups, delete lines 84-93 and replace it with the following code: cdep_groups <- 1:(J+K)
make_multivariate_crossval_results.R	27	Change "US" to desired analysis name

Script Name	Line Number(s)	Change
solvex_US.R	21	You may change the seed to reproduce results by changing the number within "set.seed()"
	23	You may change the value of the offset from x using Weibull mixture fit
	25	Change "US" to desired analysis name
	32-34	You may change the number of folds by changing the value of "k" or the number of iterations by changing the value of "maxit"
	50-52	You may change the number of folds by changing the value of "k" or the number of iterations by changing the value of "maxit"
make_publication_results.R		This script was generated solely to use for making Figures and Table information for the main article (Stull *et al.*, 2021) and Supplemental Information. If you wish to generate the same types of figures, feel free to edit the script to your needs. Please also refer to documentation on the following functions: - *vis_cont_fit* for visualizing univariate continuous fits on x - *vis_ord_fit* for visualizing univariate ordinal fits on x - *plot_x_posterior* for visualizing multivariate fits on x

3 Citations

Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Xie Y, Allen J, McPherson J, Dipert A, Borges B (2021). shiny: Web Application Framework for R. R package version 1.6.0. https://CRAN.R-project.org/package=shiny

Jantz RL, Ousley SD. FORDISC 3.0: Computerized forensic discriminant functions. Version 3.1. Knoxville, TN: The University of Tennessee, Knoxville, 2005.

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL: http://www.rstudio.com/.