Instructions for using the MATLAB program "pdPeak"

(drafted Sept. 29, 2021)

by Tomohiko Sasaki

The pdPeak method is detailed in:
Sasaki et al., Estimating sexual size dimorphism in fossil species from posterior probability densities. *Proc. Natl. Acad. Sci. U.S.A.* (2021) and Supplementary Information.

The current version of the pdPeak program is v1.0.0.

This program was developed by Tomohiko Sasaki for estimating sexual dimorphism from metric data without prior sex information (as is usually the case with fossil samples). Sexual dimorphism is here defined as the ratio of male to female population means (m/f ratio). The pdPeak method determines the point at which the posterior probability density of the parameter of interest is highest, and this program provides the estimate of the m/f ratio. It also provides the estimate of within-sex coefficient of variation (wsxCV) in the same manner. Credible intervals are determined for both estimates based on the posterior probability distributions.

All input data will be log-transformed in the program. In the log-scale dataspace, a model of two normal distributions (each for male and female) with a common variance mixed in equal proportion is assumed, and the posterior probabilities are calculated given the input data.

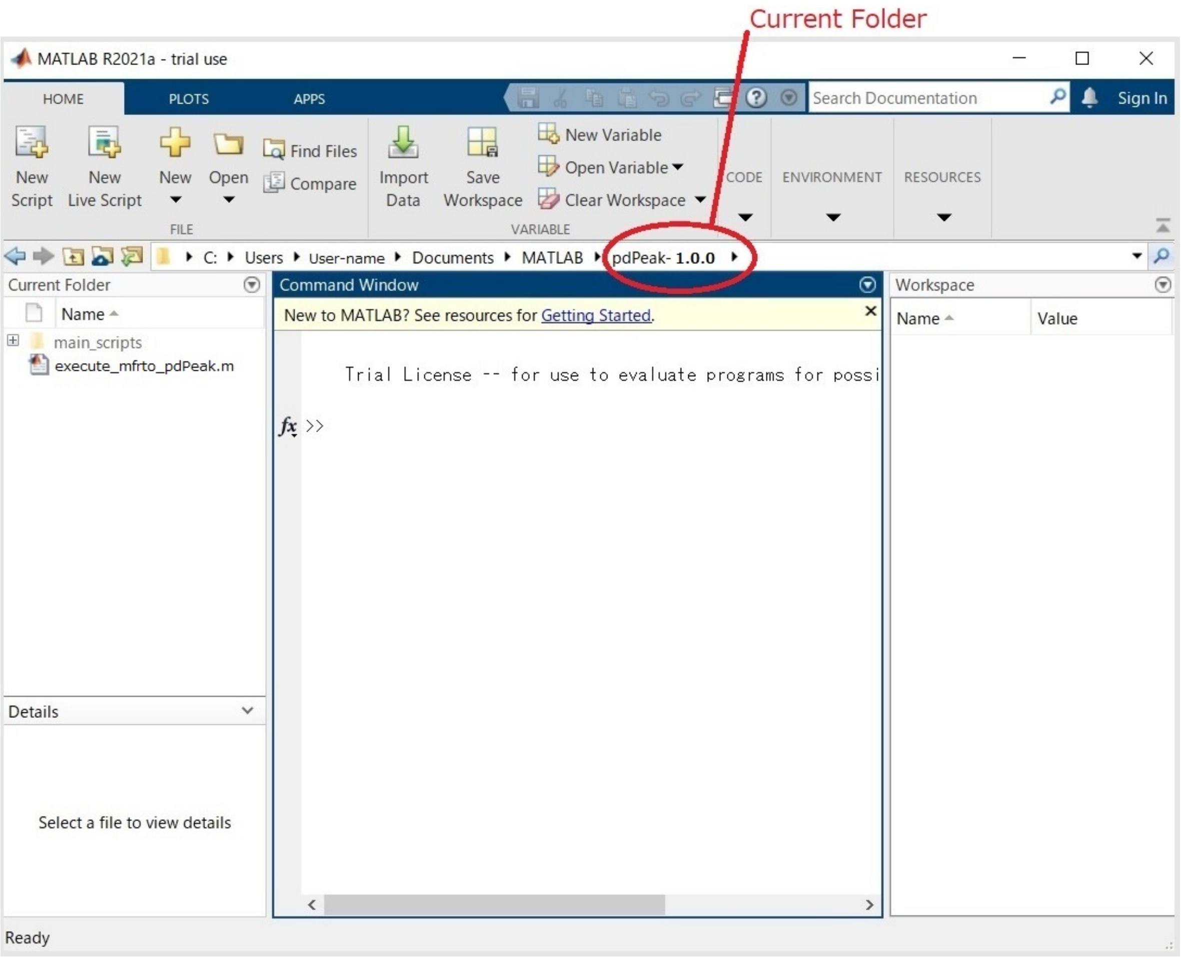
Multi-variate data are not supported.

The version of MATLAB is to be R2019b or later. No toolbox is needed.

Depending on the PC spec, especially on RAM availability, error message relating to shortage of memory may appear with risk of freezing, and the process may halt. Although not requisite, it has been observed that the program works well with a PC with RAM of 16.0 GB and CPU of Intel Core i7-6820HQ processor.

The main MATLAB code is named “execute-mfrto-pdPeak.m” which integrates all necessary other codes (which are aggregated within the folder "main scripts") and outputs the below.

1. Download the zip file (https://github.com/sxdm/pdPeak/archive/refs/tags/v1.0.0.zip) and unzip. Open MATLAB and move the entire unzipped folder, whose default name is "pdPeak-xxx" (xxx is the version name), to whatever place convenient (for example under the "MATLAB" directory). Set the Current Folder of MATLAB window to "pdPeak-xxx," then the window should look some like the one below.



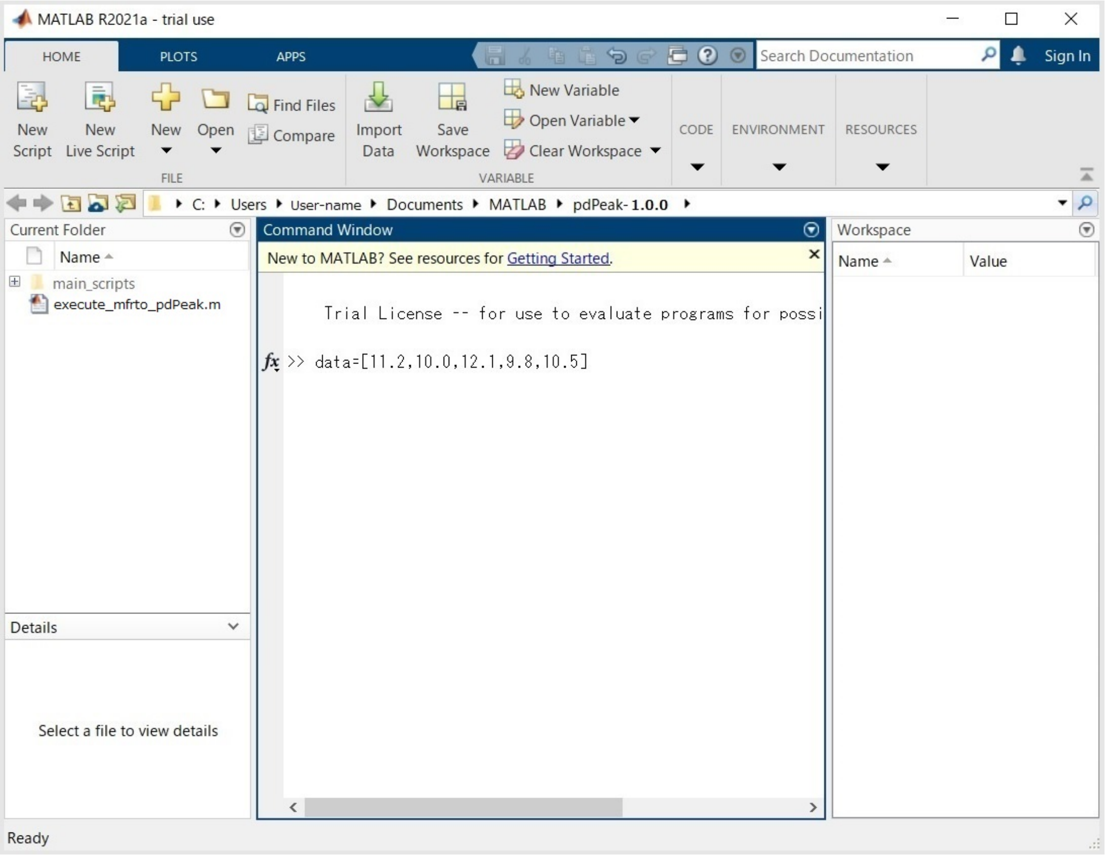
New to MATLAB? See resources for Getting Started.

```
  Trial License -- for use to evaluate programs for possi

fx >>
```

2. Make a vector of your metric data and store it in a variable named "data" (the user does not need to log-transform the data, the program does). Then, save the workspace as "metricData.mat" in the current folder.

> 2-1. For example, a row vector that contains five elements [11.2, 10.0, 12.1, 9.8, 10.5] is created by typing them in Command Window following "data = " as shown below. The elements are to be separated by commas (or spaces) and enclosed by brackets.



New to MATLAB? See resources for Getting Started.

```
  Trial License -- for use to evaluate programs for possi

fx >> data=[11.2,10.0,12.1,9.8,10.5]
```

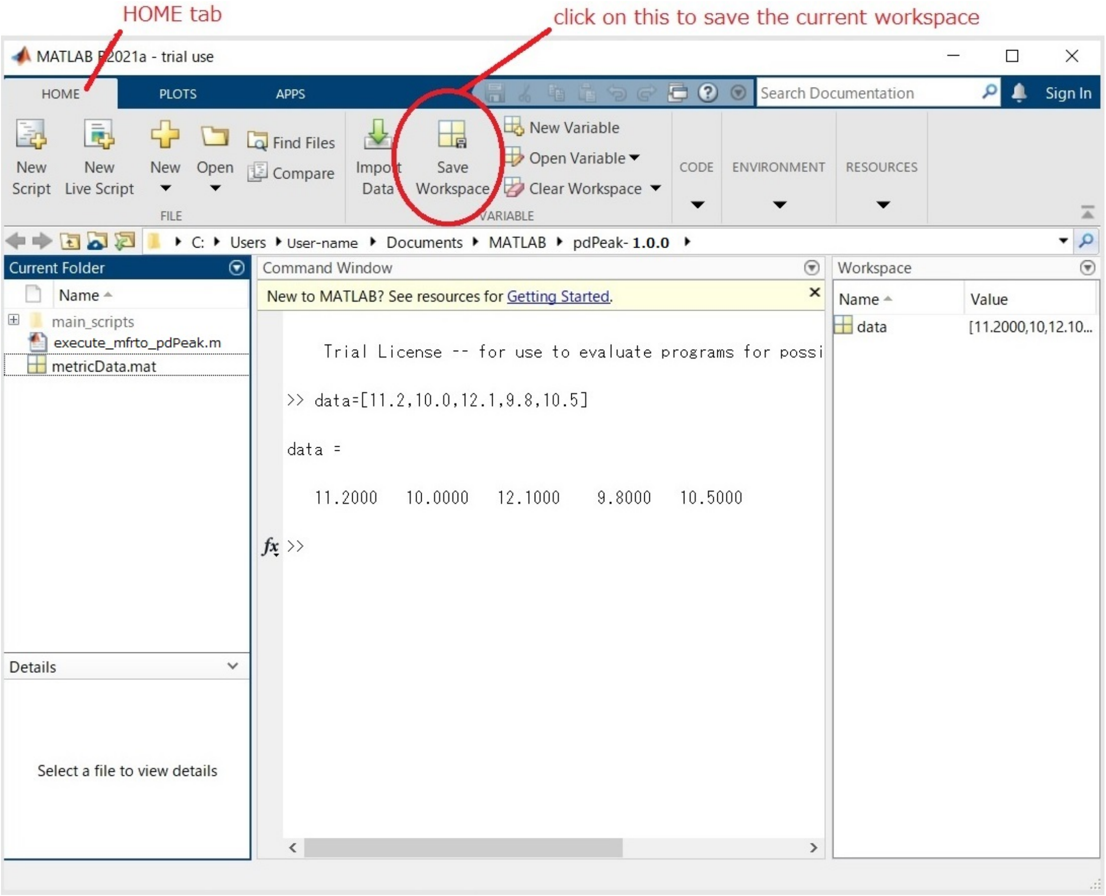
Select a file to view details

Ready

Tip: Elements can be separated also by semicolons or line breaks, in which case a column vector is created. Either a row or column vector is ok for this program. Elements within the brackets can be copy-pasted from an excel sheet.

2-2. Pressing the Enter key, a variable "data" will be created in Workspace.

2-3. Click on "Save Workspace" in the HOME tab and save the current workspace with a name "metricData.mat" in the current folder (i.e., in "pdPeak-xxx").



HOME tab

click on this to save the current workspace

```
   Trial License -- for use to evaluate programs for possi

>> data=[11.2,10.0,12.1,9.8,10.5]

data =

   11.2000   10.0000   12.1000    9.8000   10.5000

fx >>
```

3. Execute the file "execute-mfrto-pdPeak.m", then the program runs. If length of the data is >=10, the program automatically uses MCMC sampling, otherwise direct numerical calculations. The length of data has to be >=4 in any case. The process may take more than a few minutes.

3-1. To execute, double-click the icon "execute-mfrto-pdPeak.m" and then click on "Run" in the "EDITOR" tab. Or, just type "execute-mfrto-pdPeak" in Command Window and press the Enter key.

Tip: You can halt the program by pressing Ctrl+C if necessary.

4. The pdPeak estimates and other statistics will be displayed in the command window. Several figures will also pop up. Below are explanations for these display and figures.

----in the Command Window display----

"m/f ratio (by pdPeak)" shows the pdPeak estimate of m/f ratio and the following two lines show its 68% and 95% credible intervals. Credible intervals defined here is the highest density interval.

"w-sx CV (by pdPeak)" shows the pdPeak estimate of within-sex CV and the following two lines show its 68% and 95% credible intervals as defined above.

"by mean method" shows the estimate of m/f ratio by the mean method (Godfrey et al., 1993, *Am. J. Phys. Anthropol.* 90: 315-334).

"by BDI method" shows the estimate of m/f ratio by the binomial dimorphism index method (Lovejoy et al., 1989, in *Hominidae*, Jaka Book, 103-108).

"by MOM method" shows the estimates of m/f ratio and within-sex CV by the method of moment (Josephson et al., 1996, *Am. J. Phys. Anthropol.* 100: 191-206). It returns here "N/A" when the solution is imaginary.

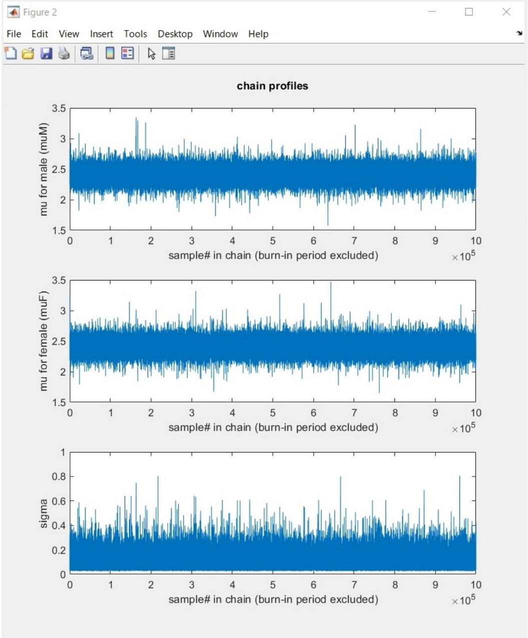
"by CVM(Plavcan1994)" shows the estimate of m/f ratio by the CV method (Plavcan, 1994, *Am. J. Phys. Anthropol.* 94: 465-476).

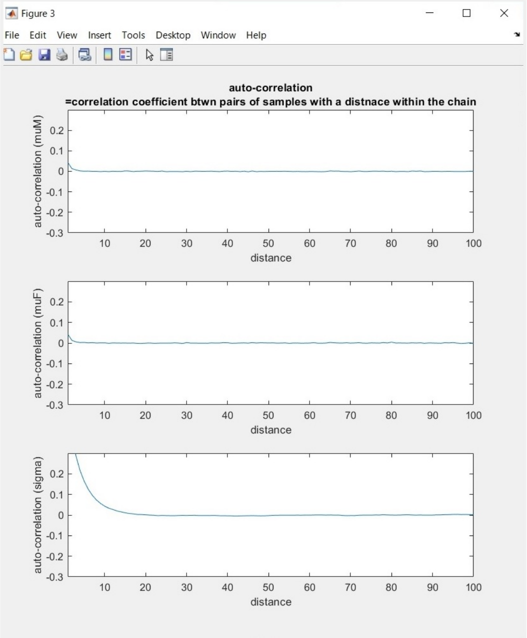
"Sample statistics" shows the basic statistics of the data including the sample size, mean, standard deviation, total CV, total CV with correction (Sokal and Braumann, 1980, *Syst. Zool.* 29: 50-66), skewness of the log-transformed data, *p*-value of the skewness upon normal null hypothesis (D'Agostino, 1970, *Biometrika* 57: 679-681).

----figures----

Data plot: This figure shows the distribution of the metric data. The center of each circle is at its value (x-axis). Circles have an automatically-determined radius so that their accumulation have a semi-histogram appearance.

MCMC (chain profiles and auto-correlation): Two figures, as shown below, pop up when the MCMC algorithm is used in the process. MCMC is used to approximate the posterior probability distributions while avoiding high computational loads when the data length (number of specimens) is relatively large. The algorithm generates a chain of 1,000,000 MCMC samples (excluding an initial short phase called the burn-in period) that should conform to the posterior distribution as a whole. Sufficiency of the chain length may be qualitatively evaluated by referring to the following two figures (sets of plots).





The left panel shows the chain profiles of the model parameters, i.e., male mean, female mean, and common sd (all in log-scale), from top to bottom. All the plots should not show stagnations nor excessive (outlying) spikes throughout the whole chain length. The above is an ok example, while the plots shown in the below figure is a bad example that shows multiple outlying spikes. The latter indicates that the range of the posterior distribution is not sufficiently explored by the MCMC chain.

The right panel shows the auto-correlation graphs for the same parameters as in the left panel plots. These graphs show correlation coefficients (y-axis) between pairs of MCMC samples within the chain at certain distances (x-axis). The sample correlation within the chain should decrease as the samples are distantly separated. Near-zero correlation at a short distance is an indication of good independence of the samples within the chain. Here, the plots show the correlations with distances as large as

100. It is preferable that the correlation is negligible before the distance becomes 100. Again, the above figure is a good example and the below figure is a bad example.

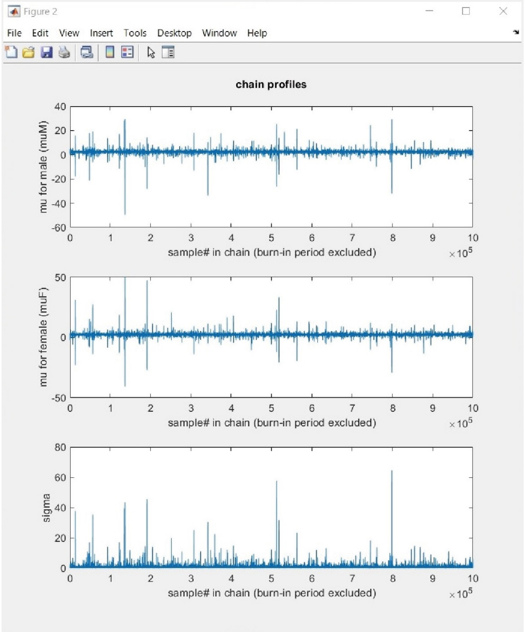
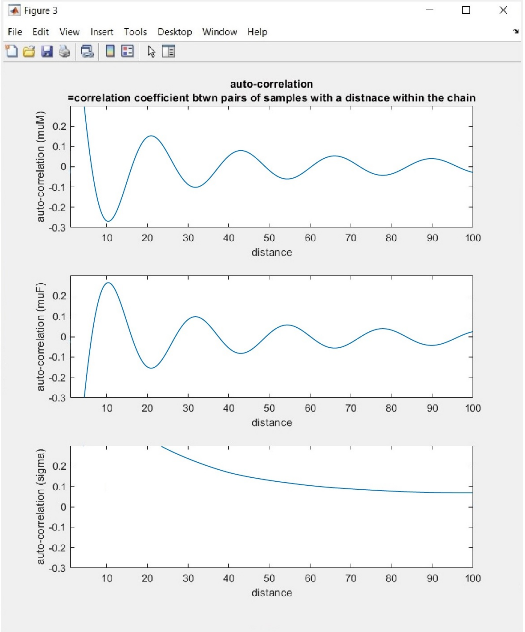


Figure 2
File Edit View Insert Tools Desktop Window Help
chain profiles
mu for male (muM)
40
20
0
-20
-40
-60
0 1 2 3 4 5 6 7 8 9 10
sample# in chain (burn-in period excluded)
×10^5
mu for female (muF)
50
0
-50
0 1 2 3 4 5 6 7 8 9 10
sample# in chain (burn-in period excluded)
×10^5
sigma
80
60
40
20
0
0 1 2 3 4 5 6 7 8 9 10
sample# in chain (burn-in period excluded)
×10^5



Empirically, however, when data length is >= 10 (the default setting), the default chain length (1,000,000) should be sufficient.

Sexual dimorphism (m/f ratio): This figure is for the assessment of the posterior probability distribution of the m/f ratio. When the MCMC is used, the histogram of the MCMC samples is shown. The green line is the density curve estimated from the histogram. When the MCMC is not used, the density curve calculated via numerical computation is shown. The red diamond is at the pdPeak estimate. The dashed or dotted horizontal lines indicate the density levels at 68% or 95% credible intervals, respectively. Note that the density curve is that of the logarithm of the m/f ratio (i.e., $\mu_m - \mu_f$, where μ_m and μ_f are the log-scale male and female means, respectively). Thus, the x-axis labels here are not equidistant (see Sasaki et al. 2021 SI Appendix Text for details).

Within-sex CV: This figure is for the assessment of the posterior probability distribution of within-sex CV (wsxCV). When the MCMC is used, the histogram of the MCMC samples is shown. The green line is the density curve estimated from the histogram. When the MCMC is not used, the density curve calculated via numerical computation is shown. The red diamond is at the pdPeak estimate. The dashed or dotted horizontal lines indicate the density levels at 68% or 95% credible intervals, respectively. Note that, the density curve is that of $\sqrt{\ln(\mathrm{wsxCV}^2 + 1)}$ (i.e., σ, the within-sex standard deviation in log scale), and thus the x-axis labels here are not equidistant (see Sasaki et al. 2021 SI Appendix Text for details). However, these are nearly equidistant in most cases (when wsxCV<<1).

Bivariate density plot: This figure is for the assessment of the bivariate posterior probability of within-sex CV (x-axis) and m/f ratio (y-axis). When the MCMC is used, the densities indicate the frequencies of the MCMC samples. When the MCMC is not used, the numerically calculated densities are shown. Dotted contour lines show theoretical combined-sex CV levels, which is a function of the m/f ratio and within-sex CV under the model assumption. Note that, the density is that of the logarithm of the m/f ratio and $\sqrt{\ln(\mathrm{wsxCV}^2 + 1)}$ (see above). Thus, the y- and x-axis labels here are not equidistant, although the x-axis is in nearly equidistant units in most cases (when wsxCV<<1) (see Sasaki et al. 2021 SI Appendix Text for details).