NMπ (ver. 2.0) – The User Manual

Igor J. Chybicki

Department of Genetics, Kazimierz Wielki University
Bydgoszcz, Poland

Last update: 17-02-2023

Contents

1 Brief introduction

NMπ is a successor of the NM+ software (Chybicki and Burczyk 2010), implementing most of its best features. However, NMπ offers more flexible data treatment, including explicit model for separate sexes, possibility of mixing dispersed (seedlings) and non-dispersed (seeds) data, cytotype information usage, biparental phenotypic differentials and the other minor improvements. Starting from version 2.0, NMπ implements the Bayesian approach designed to avoid false positive selection gradients due to omitting determinants of reproductive success (Chybicki et al. 2021, Chybicki, in press). The Bayesian approach, called the hierarchical neighborhood model, is fully functional for both seeds and seedlings.

2 What can NMπ do?

Basically, NMπ is useful in characterizing plant mating patterns. The analysis is based on a parentage model called the neighborhood model which explicitly accounts for dispersal kernels, self-fertilization, effects of phenotypic characters on reproductive success and genotyping problems. A list of possible model parameters include:

- frequency of self-fertilization
- frequency of seed and pollen immigration from surrounding sources
- mean distance of seed and pollen dispersal
- shape of seed and pollen dispersal kernels (probability models)
- directionality (anisotropy) rate for seed and pollen dispersal kernel
- prevailing direction (azimuth) of seed and pollen dispersal
- effects of phenotypic characters on reproductive success (selection gradients)
- single-locus genotyping error rates

The full list of parameters is given in Table 1.

In addition, as a result of the analysis, NMπ generates file containing inferred genealogies (parentages). Genealogies are fully based on the neighborhood model. It means that the most likely mother and father are inferred accounting for non-random seed and pollen dispersal as well as (if included) selection gradients.

Furthermore, mother or both parents can be assumed as known a priori for individual progeny. Each progeny individual can be also assumed to represent dispersed or non-dispersed individual. Non-dispersed individuals are assumed to be collected from individual plants (like seeds) and, consequently, they provide no information on seed dispersal and maternal reproductive success. On the contrary, dispersed individuals are assumed to be collected as independent plants (possibly seedlings or dispersed seeds), for which coordinates were specifically measured. Dispersed individuals provide information about both female and male reproductive success as well as seed and pollen dispersal.

Table 1. List of parameters used in the neighborhood model

Parameter	Description	Comments
ms	frequency of seed immigration from surrounding seed sources	naturally between 0 and 1
s	frequency of self-fertilization	naturally between 0 and 1
mp	frequency of pollen immigration from surrounding pollen sources	naturally between 0 and 1
ds	mean distance of seed dispersal (mean of forward dispersal kernel)	naturally positive
dp	mean distance of pollen dispersal (mean of forward dispersal kernel)	naturally positive
bs	Shape parameter of seed dispersal kernel (exponential-power of Weibull)	must be positive
bp	Shape parameter of pollen dispersal kernel (exponential-power of Weibull)	must be positive
ks	Intensity (rate) of directionality (anisotropy) in seed dispersal; (rate parameter in von Misses distribution)	interpretation as the slope in regression
kp	Intensity (rate) of directionality (anisotropy) in pollen dispersal; (rate parameter in von Misses distribution)	interpretation as the slope in regression
as	Prevailing direction (azimuth) of seed dispersal; (dominant angle in von Misses distribution)	expressed as %2π; zero is North; the value of 0.5 is pi (in radians) or 180 (in degrees) – South
ap	Prevailing direction (azimuth) of pollen dispersal; (dominant angle in von Misses distribution)	expressed as %2π; zero is North; the value of 0.5 is pi (in radians) or 180 (in degrees) - South
mtyp	frequency of mistyping of an allele (wrong allele calling)	naturally between 0 and 1; locus-specific
mcytyp	frequency of mistyping of a cytotype (wrong haplotype calling)	naturally between 0 and 1; locus-specific
g_k	selection gradient; effect (rate) of k-th phenotypic character on female reproductive success	regression parameter (as the slope in soft-max regression)
b_k	selection gradient; effect (rate) of k-th phenotypic character on male reproductive success	regression parameter (as the slope in soft-max regression)
bipar_k	indicator modifying the meaning of male selection gradient for k-th phenotypic character	not estimable; if 0 - male selection gradient is the slope of the effect of k-th character on male reproductive success, if 1 - male selection gradient is the slope of the effect of distance (absolute difference) between values of k-th character measured for female and male
maternal_cy	probability of maternal transmission of cytotype (cp or mtDNA)	can be set constant or estimable; 1 means that cytotype is transmitted from mother, 0 – from father; 0.5 – with equal prob. from mother or father

neighborhood	maximum distance between individuals included into considerations about parentage; radius of the neighborhood	this parameter influences data treatment; likelihood functions between different neighborhoods are generally incomparable
log.fem.fec.i	Logarithm of female fecundity for the i-th adult	this parameter is estimated if the Bayesian approach is chosen
log.fem.mal.i	Logarithm of male fecundity for the i-th adult	this parameter is estimated if the Bayesian approach is chosen
Posterior frequency of model.k	Posterior frequency of the k-th regression model for fecundity	this parameter is estimated if the Bayesian approach is chosen; it can be treated as the posterior estimate of the probability for the k-th model
Posterior frequency of var.n	Posterior frequency of the n-th explanatory variable in the regression model for fecundity	this parameter is estimated if the Bayesian approach is chosen; it can be treated as the posterior estimate of the inclusion probability for the n-th variable
sigm.g	Standard deviation of the normal distribution for log female fecundity	this parameter is estimated if the Bayesian approach is chosen
sigm.b	Standard deviation of the normal distribution for log male fecundity	this parameter is estimated if the Bayesian approach is chosen
R^2.g	Bayesian analog of the determination coefficient estimated for the best regression model for female fecundity	this parameter is estimated if the Bayesian approach is chosen; it reflects the proportion of variance in individual female fecundity explained by the best regression model
R^2.b	Bayesian analog of the determination coefficient estimated for the best regression model for male fecundity	this parameter is estimated if the Bayesian approach is chosen; it reflects the proportion of variance in individual male fecundity explained by the best regression model

3 How to prepare data file?

Data file is a standard text file (so-called 'flat file'). It is recommended to prepare a tab-delimited text (but space-delimited text is also okay), so a spreadsheet program (e.g. Excel or Calc) can be convenient for this purpose. The first line in the data file is the header of the following structure:

```
np  no   nl   nf
```

where np – a number of (candidate) parents in the file, no – a number of progeny in the file, nl – number of loci (nuclear genetic markers only), nf – a number of phenotypic characters in the file. Second and further lines contain information about individuals. Consequently, NMπ expects the following structure:

```
np no  nl  nf
Line for parent #1
Line for parent #2
:
Line for parent #np
Line for progeny #1
Line for progeny #2
:
Line for progeny #no
```

Each line containing data for a parent must start with 0 (zero generation). Then, the ID number, X coordinate, Y coordinate, cytotype, genotype, phenotypic characters and femaleness index follows. For 3 loci and 2 phenotypic characters the example parental line looks like this:

```
0   13   -9.1 10.1 4    122  122  213  217  -1   -1   -0.93 1.11    1
```

In the above example:

- 13 is the ID number of the individual. Must be positive integer.
- -9.1 and 10.1 are X and Y coordinates. Must be real.
- 4 is a cytotype. Must be integer. If mtDNA or cpDNA was not assayed, all cytotypes must be equal. However, if cytotypes were generally assayed but some individuals failed to be genotyped, put -1 (missing cytotype). Note that the program requires this column even if information on cytotypes is unavailable. In this case one can put either '-1' (missing data) or the same integer for individuals (as if all individuals have the same cytotype).
- Genotype is 122/122, 213/217, missing/missing. Every single allele must be integer. Missing genotype is denoted by double -1 (tab-separated!).
- Phenotypes are -0.93 and 1.11. It is recommended to use standardized measurements of phenotypic traits. Then, zero can be used for missing data. If number of phenotypic characters is set to zero (`nf`=0) then genotypes are followed directly by femaleness (next value).
- Femaleness is 1. Note that femaleness = 1 denotes 100% female. If femaleness = 0 then an individual is 100% male. Femaleness of 0.5 means either that an individual is a hermaphrodite or undetermined (unobserved sex/missing data).

Each line containing data for an offspring must start with 1 (first generation). Then, the ID number, X coordinate, Y coordinate, cytotype, genotype, mother, father and dispersal index follows. For 3 loci the example offspring line looks like this:

```
1   111  -0.1 -0.9 2    122  128  213  213  110  114  13   -1   0
```

In the above example:

- 111 is the ID number of the individual. Must be positive integer.
- -0.1 and -0.9 are X and Y coordinates. Must be real. Note that for offspring collected from mother plants (non-dispersed seeds), X and Y should be equal to maternal coordinates. However, such individuals do not provide information on seed dispersal, so their

coordinates are not used in practice. XY coordinates must be in a Cartesian coordinate system.

- 2 is a cytotype. Must be integer. If mtDNA or cpDNA was not assayed, all cytotypes must be equal.
- Genotype is 122/128, 213/213, 110/114. Every single allele must be integer. Missing genotype is denoted by double -1 (tab-separated!).
- Maternal individual is 13. This number is equal to the ID number of parental individual in the file. If mother is not known a prior then -1 is used.
- Paternal individual is -1. -1 is used when father is not known a priori. Note that NMπ does not allow for unknown mother and known father.
- 0 at the end means that the offspring is non-dispersed. Note that if an offspring in non-dispersed then NMπ expects mother to be known a priori (so -1 for mother generates an error). However, an offspring with a known mother can be assumed to represent dispersed individual (e.g. dispersed seed for which mother is known based on pericarp genotyping). In the case of dispersed offspring the last value in the line should be 1.

-

Important note:
Note that the seed immigration rate, seed dispersal kernel parameters and selection gradients for female reproductive success are estimated based on dispersed progeny only. For example, `ms` parameter reflects the estimated proportion of dispersed progeny being a result of seed immigration among all dispersed progeny in the sample. Consequently, if dispersed and non-dispersed data are mixed together then `ms` informs about the proportion of seed immigrants among a subsample of dispersed progeny.

4 How to prepare input files?

NMπ requires four appropriately formatted text files, i.e. the information file (obligatory named 'info.txt'), the main data file (named freely), the file containing genotyping error rates (named freely) and the file containing starting values for selection gradients (selection gradients are effects of phenotypic characters on a reproductive success).

Important note:
When using a command line version, all input files must be prepared by the user and saved in the same folder as the NMpi.exe. On the other hand, when using a graphic user interface (GUI) for Windows (wNMpi.exe), input files are created and managed by the GUI and data file is the only file that must be prepared by the user. Moreover, when GUI is used to control NMπ, data file can be placed in a freely chosen folder, which will be then used as a working directory for NMpi.exe.

4.1 Information file

The example information file (info.txt) looks as follows:

```
0.050    0    0.000     1.000
0.010    0    0.000     1.000
0.500    0    0.000     1.000
2.708    0    0.000     9.210
3.219    0    0.000     9.210
1.000    0    0.050     3.000
1.000    0    0.050     3.000
0.000    0
0.000    0
0.000    0
0.000    0
1   1    1
0.500
-100
20  50
1
0.000606
0
EXAMPLE.txt
EXAMPLE.out
EXAMPLE.err
EXAMPLE.sel
EXAMPLE.par
```

Lines 1-7 contain 4 values each: `val, incl, min, max`

`val` is an initial value of a parameter (must be real number)

`incl` is a binary (0 or 1) indicator of whether a parameter is estimated (1) or constant (0) (must be integer)

`min, max` are min. and max. value for a parameter (both must be real numbers).

- Line 1 is for frequency of seed immigration (`ms`)
- Line 2 is for frequency of self-fertilization (`s`)
- Line 3 is for frequency of pollen immigration (`mp`)
- Line 4 is for mean distance of seed dispersal (`ds`)
- Line 5 is for mean distance of pollen dispersal (`dp`)
- Line 6 is for shape parameter of seed dispersal kernel (`bs`)
- Line 7 is for shape parameter of pollen dispersal kernel (`bp`).

Lines 8-11 contain two values each: `val, incl`

`val` is an initial value of a parameter (must be real number)

`incl` is a binary (0 or 1) indicator of whether a parameter is estimated (1) or constant (0) (must be integer)

- Line 8 is for rate of directionality (anisotropy) of seed dispersal kernel (`ks`)
- Line 9 is for rate of directionality (anisotropy) of pollen dispersal kernel (`kp`)
- Line 10 is for prevailing direction (azimuth) of seed dispersal (`as`)
- Line 11 is for prevailing direction (azimuth) of pollen dispersal (`ap`).

Important note:
When `incl` is set to 1, a given parameter is set estimable. However, in this case the initial value of this parameter cannot be zero!

Line 12 contains 3 numbers: `disp_typ, seed_kernel, pollen_kernel`
`disp_typ` is a binary (0 or 1) indicator of whether reciprocals of seed and pollen mean dispersal distance (1/`ds` and 1/`dp`) are used as parameters (1) or not (0). Zero means that mean dispersal distances are estimated. Note that confidence bounds for `ds` and `dp` are computed only if `disp_typ` = 1.
`seed_kernel` is an integer indicator of kernel function for seed dispersal: 1 = exponential-power, 2 = Weibull, 3 = Tufto, 4 = Log-normal, 5 = Power-law
`pollen_kernel` is an integer indicator of kernel function for pollen: 1 = exponential-power, 2 = Weibull, 3 = Tufto, 4 = Log-normal, 5 = Power-law and 0 = linked to seed dispersal (seed and pollen dispersal is modeled using a single function; can be useful to test whether seed and pollen dispersal kernels differ or not)

Line 13 contains one number: `maternal_cy`
`maternal_cy` is a probability of maternal transmission of cytoplasmic DNA (must be real between 0 and 1)

Line 14 contains one number: `neighborhood`
`neighborhood` equals the maximum distance between individuals to be considered as neighbors; when maternity and/or paternity is reconstructed only neighbors are taken into consideration. Use any negative value of `neighborhood` in order to set unlimited distance. In this case, all possible configurations (pairs or trios) of individuals will be taken into consideration.

Line 15 contains two numbers: `max_piter` and `max_iter`
`max_piter` is a maximum number of iterations for initial optimization of parameter values (must be positive integer to enable initial optimization procedure)
`max_iter` is a maximum number of iterations; This parameter is used to choose between the maximum likelihood estimation under the classec neighborhood model (if `max_iter` is a positive integer) or the Bayesian estimation under the hierarchical neighborhood model (if `max_iter` is a negative integer). In the latter case, the parameter denotes the number of MCMC iterations and should be large (e.g. -100000 informs that the Bayesian estimation is chosen with 100,000 MCMC iterations).

Important note:
The Bayesian estimation under the Hierarchical Neihborhood Model is chosen when `max_piter` and `max_iter` are set to a negative integer. In this case, the absolute value of `max_iter` is equal to the number of MCMC sampling cycles.

Line 16 contains one number: `nth`
`nth` is a number of threads; number of processors used for processing (must be positive integer)

Line 17 contains one number: `mep`
`mep` is a machine epsilon (default value of 0.000606 is likely the best) (must be positive real)

Line 18 contains a single number: `writecov`
`writecov` is a binary (0/1) integer indicator of whether NMpi creates an additional output file (0 = the file is not created, 1 = the file is created). If the Maximum Likelihood estimation is chosen (see `max_iter` parameter), the additional output file contains the Variance-Covariance matrix. The matrix can be helpful if confidence bounds around dispersal kernel parameters are needed. More information is provided in the section 'Estimation strategies'. If the Bayesian estimation is chosen, the additional output file contains the full MCMC output, i.e. a series of parameter values across the sampling stage of MCMC iterations.

Lines 19-23 contains file names of the following order:
Line 19: name of the data file (existing input file)
Line 20: name of the results file (output file created by NMpi)
Line 21: name of file containing settings for error rates (existing input file)
Line 22: name of file containing settings for selection gradients (existing input file)
Line 23: name of file containing settings for error rates (output file created by NMpi)
(File names must be given in separate lines!)

4.2 File containing typing error settings

NMπ offers the same model as implemented in NM+ (see the manual of NM+). Basically, the model accounts for random mistyping of alleles. In this way all possible mistakes/mutations are approximately covered.

The file containing typing error settings for 3 loci in the data file has the following structure (values given as an example):

```
0   0
0.000
0.002
0.078
0.000
```

The file begins with the header line. The header contains two numbers: `incl_mtyp` and `incl_mcytyp` – both are binary indicators equal 1 if error rates are treated as estimable parameters or 0, otherwise (must be integer 0 or 1). However, `incl_mtyp` is for nuclear markers and `incl_mcytyp` is for a cytoplasmic marker. Then, there are as many rows as many loci declared in the data file + 1 (e.g. there must be 9 lines for 8 loci). Each line contains a single value of `mtyp`, a frequency of allele mistyping for a given locus. The last line contains a frequency of cytotype mistyping (`mcytyp`). `mtyp` and `mcytyp` must be real numbers between 0 and 1. Note that a value of zero will result treating `mtyp` or `mcytyp` as a constant value (i.e. not updated during estimation).

Important note:
`mtyp` can be set to zero for some loci and any value between 0 and 1 for another loci. If `incl_mtyp` is set to 1 (meaning that error rates are treated as estimable parameters) NMπ will then treat error rates depending on their values either as constants (when equal to zero) or as estimable quantities (when equal to a non-zero value).

4.3 File containing selection gradients

The file containing selection gradient settings for 2 phenotypic characters in the data file has the following structure (values given as an example):

```
0.000   0    0.000    0    0
0.000   0    0.052    1    0
```

There are as many rows as many phenotypic characters declared in the data file. Each line contains five values: `val_f`, `incl_f`, `val_m`, `incl_m` and `bipar_m`.

- `val_f` is a selection gradient for a given phenotypic character related with female reproductive success
- `incl_f` is a binary indicator of whether a given parameter is estimated (1) or constant (0) (must be integer)
- `val_m` is a selection gradient for a given phenotypic character related with male reproductive success
- `incl_m` is a binary indicator of whether a given parameter is estimated (1) or constant (0) (must be integer)
- `bipar_m` is a binary indicator of whether a male fitness is a function of a value of a given phenotypic character (0) or is a function of an absolute difference between a value of a given phenotypic character measured for maternal and paternal individuals (1) (must be integer); one example when `bipar_m=1` would be useful is consideration of flowering phenology as a putative factor of mating success. If flowering dates are available then, rather than simple flowering date of a candidate father, the difference between flowering dates for a mother and a father is likely the actual predictor of mating success. Generally, `bipar_m` can be useful if one wishes to verify a hypothesis about assortative/disassortative mating, when phenotypically similar individuals mate more frequently (less frequently) than at random.

Important note:
Initial values (`val_f` and `val_m`) in the file containing selection gradients should be set differently for the Maximum Likelihood (MLE) estimation and the Bayesian estimation (BE). In the case of MLE, initial values must be non-zero for phenotypic characters having binary indicators (`incl_f` or `incl_m`) set to 1. In the case of BE, all initial values (`val_f` and `val_m`) must be exactly zero regardless of binary indicator settings.

5 How to interpret output files?

If the Maximum Likelihood estimation is chosen, NMπ generates up to four output files containing the results and three additional output files which help in continuing analysis with current estimates taken as starting points. If the Bayesian estimation is chosen, NMπ generates a main file containing results summary as well as the file with parentage assignments. Optionally, NMπ generates an output file containing a series of parameter updates across the MCMC iterations (with the default frequency of every 20 iterations).

5.1 Output files containing results

5.1.1 The Maximum Likelihood analysis under the standard Neighborhood model

If the maximum value of log-likelihood of the model is found (or maximum number of iterations is reached), NMπ writes the results to output files using names declared in the info.txt file. See the structure of info.txt file (section 4.1) for details.

The main output file contains estimates of the all parameters in the neighborhood model as well as the maximum value of log-likelihood of the model is given (`Final log-L`). The structure of the main output file is rather self-explanatory. Only, 'Estim.:' rows contain estimates of parameters and 'SE:' rows contain standard errors approximated with the Hessian. If a given parameter was treated as constant (not estimated) then SE value equals -1. Also, a value of -1 can incidentally be an effect of bad properties of the Hessian.

Second output file contains inferred genealogies for progeny, given the values of parameters of the neighborhood model. Structure of this file is as follows:

```
Prog        Mo1        Fa1         Pr1        Mo2        Fa2         Pr2
111           1         -1      0.9375          1        201      0.0554
```

The first column contains offspring ID number. Then, the most likely genealogy is given together with the (a posteriori) probability: Mo1 - the most likely mother (ID number), Fa1 – the most likely father (ID number). If ID number equals -1 then the most likely parent is outside the study population. Last three columns show the second most likely genealogy using the same structure.

In the example above, two the most likely genealogies for progeny '111' are {1,-1} and {1,201}. The first genealogy has assigned the probability of 0.9375. This means that, in the face of the observed genotypes and coordinates (and possibly additional data, such as sex, phenotypes and cyDNA), given the most likely estimates of the neighborhood model, adult '1' and an unknown individual from outside the sampling site are the true pair of parents for progeny '111' with the probability of 93.8%. For comparison, the second most likely genealogy {mother=1, father=201; both within the sampling site} has assigned probability of 5.5% only. Consequently, the most likely genealogy is almost 17 times more likely than the concurring scenario. In addition, any other possibility has the probability of no more than 0.7%. Consequently, the most likely genealogy is at least 130 times more likely than the third, fourth and so-on best genealogy, having a very good support overall.

In addition, two output files can be also generated. If `dist_typ` is set to 1, NMπ creates 'kernelconf.out' file containing the approximate 95% confidence interval around dispersal kernel

parameters `ds`, `dp`, `bs` and `bp`. If `writecov` is set to 1, NMπ creates 'cov.out' file containing the variance-covariance matrix for all the parameters in a model.

5.1.2 The Bayesian analysis under the Hierarchical Neighborhood model

The Bayesian estimation procedure generates a single text file with the results summary. The file has the same name as the results file (set in the info.txt file) except for the default extension of "*.hnm". The "*.hnm" file consists of several sections, each starting with the "#" mark.

Section	Content
# Model support	`LogL.avg.theta.` – the log-likelihood of the model for posterior estimates of parameters `avg.LogL` – the average log-likeihood of the model across MCMC iterations `var.LogL` – the variance of log-likelihood values across MCMC iterations
# Posterior distribution of model parameters	Posterior estimates of model parameters, i.e. posterior median and limits of the 95% credible interval (hpdl(95%) stands for lower bound and hpdh(95%) stands for higher bound) Note that the parameter list includes parameters of individual fecundity `log.fem.fec.j` and `log.mal.fec.k` for the j-th and k-th adult, respectively.
# Posterior distribution of female fecundity models (analogously for the male fecundity models)	Estimates of posterior distribution of regression models for female fecundity. Competing regression models contain different subsets of explanatory variables. The model number (M) decodes individual inclusion indicators as a single integer value according to the formula: $M = y_1 \times 2^0 + y_2 \times 2^1 + \ldots + y_W \times 2^{W-1}$ The backward translation of M into a series of W indicators follow the rule: $y_w \mid M = (M \; div \; 2^{w-1}) \; mod \; 2$ e.g. for W=3 variables, the model 5 corresponds to the following series of indicators: $y_1 \mid 5 = (5 \; div \; 2^{1-1}) mod \; 2 = 1$ $y_2 \mid 5 = (5 \; div \; 2^{2-1}) mod \; 2 = 0$ $y_3 \mid 5 = (5 \; div \; 2^{3-1}) mod \; 2 = 1$ In other words, the model 5 includes 1st and 3rd explanatory variable.
# Posterior inclusion frequency of phenotypic variables (female fec.) (analogously for the male fecundity models)	Estimates of posterior inclusion frequency of phenotypic variables in a regression model for female fecundity. The frequency can be treated as the estimate of the probability that a given variable is in the regression model.
# Regression parameter estimates for the best female fecundity model (analogously for the male fecundity models)	Estimates of selection gradients for female fecundity (`gamma.w` for the w-th variable), including posterior median and the limits of 95% credible interval In addition, the standard deviation of female log-fecundity (`sigm.g`) and the proportion of variance in

	fecundity explained by the best regression model
# Metropolis-Hastings acceptance rates	This section contains acceptance rates for proposed parameter values across MCMC iterations The value below 25% or above 45% should be treated as an indicator of bad mixing; in such case, the parameter estimates may be of bad quality

Second output file contains inferred genealogies for progeny, given the values of parameters of the neighborhood model. Structure of this file is the same as for the Maximum Likelihood estimation.

5.2 Additional output files

If the Maximum Likelihood estimation is chosen, together with output files containing results, NMπ creates three additional output files, i.e. 'info.out', 'errors.out' and 'gradients.out'. These files have the same structure as the input files described in section 4, except that they contain final estimates of all parameters. If one wishes to continue the analysis with the final parameter values from a previous run, contents of those output files can be copy-pasted into input files prepared for the next run.

When using GUI for Windows (see the next section), additional output files can be used as a source of settings (see the section 6.2).

6 How to use GUI for Windows (wNMpi)?

NMπ is written in GNU Fortran and, therefore, it lacks a fancy user interface. Windows users, however, are provided with a graphic user interface (GUI) which compensates for this disadvantage. It should be stressed that GUI (wNMpi.exe) is an interface between the user and nmpi2.exe and requires nmpi2.exe to perform any analysis.

The interface is split into tabs, each dedicated to a specific task. The tab 'Data' is for data management, the tab 'Parameters' is for settings related with estimation, the tab 'Console output' is for capturing the output of NMpi.exe when running and the tab 'Preferences' is for the various settings. Note that 'Parameters' tab is active only when data file is selected.

Data file must be created following description in 'How to prepare data file?' section. To prepare analysis, first a data file must be selected.

To run analysis just press [Run] button. To cancel running analysis press [Cancel] button. To exit GUI press [Close] button.

6.1 'Data' tab

Using 'Data' tab, a (properly formatted) data file can be selected using […] button right to 'Data file:' text box. In addition, output file name can be set. The output file name will be used to save results of estimation (file with extension '.out') and to create file with inferred genealogies (file with extension '.par').

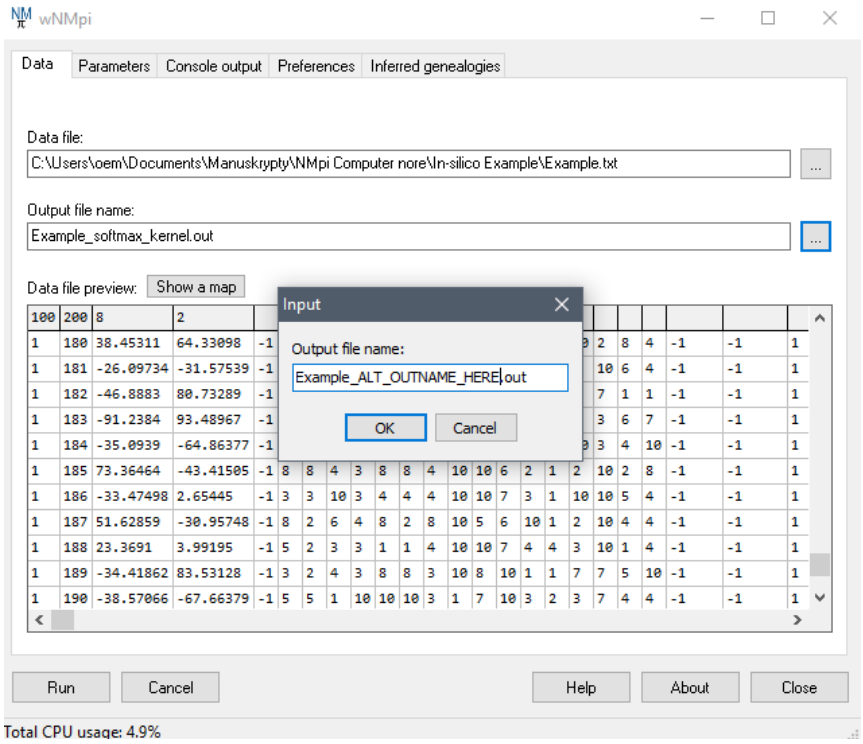

Fig. 1. […] button right to 'Output file name:' edit box allows to set new output file name. Note that this name will be used for both the main results file and the file containing inferred genealogies. However, the file containing genealogies will have extension automatically changed to '.par'. For details regarding output files see section 5.1.

Using [Show a map], one can inspect visually if coordinates of individuals were properly read and the distribution of individuals reflects the actual distribution.

6.2 'Parameters' tab

This section contains many controlling elements to set-up the model and the estimation.

Important notes:

- 'Parameters' tab becomes fully active only if data file is selected.
- After clicking [Run], all input files described in 4. are created automatically and NMpi.exe is executed with the console output shown in the 'Console output' tab.
- The important option in the 'Parameters' tab is the checkbox to switch between the standard model (the Maximum Likelihood estimation) and the hierarchical model (the Bayesian Estimation).

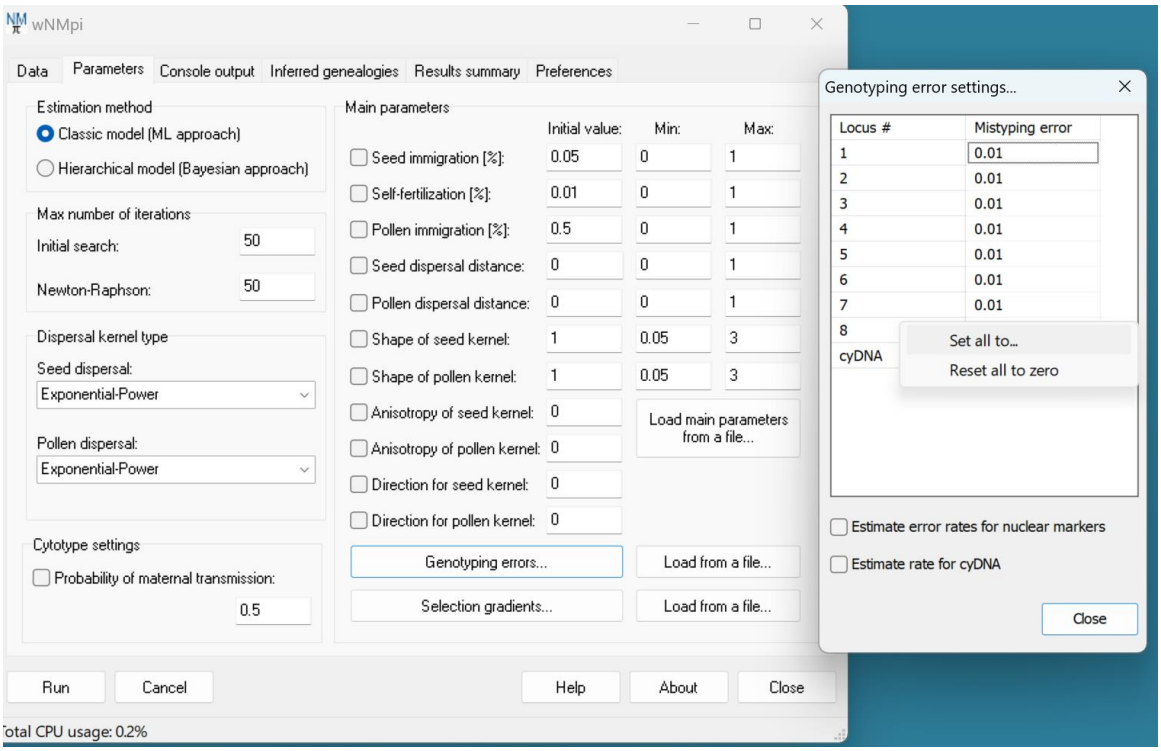

Fig. 2. When setting genotyping error rates, it is possible to set all the parameters to a given value using the right-click pop-up menu.

Main parameter settings can be loaded from a file selected using [Load main parameters from a file…] button. The file must have the same structure as the information file described in section 4.1. 'info.out' file is used as a default file name.

To edit genotyping error model settings press [Genotyping errors…] button. It opens additional window with parameters related to genotyping errors. [Load from a file…] button right to [Genotyping errors…] allows to load genotyping error parameter settings from a selected file. The file must have the same structure as the input file described in section 4.2. 'errors.out' file is used as a default file name.

To edit selection gradients settings press [Selection gradients…] button. It opens additional window with parameters related to selection gradients. [Load from a file…] button right to [Selection gradients…] allows to load selection gradient parameter settings from a selected file. The file must have the same structure as the input file described in section 4.3. 'gradients.out' file is used as a default file name.

6.3 'Console output' tab

'Console' tab shows the live output of nmpi2.exe. This tab allows the user to track the progress of the analysis.

6.4 'Inferred genealogies' tab

'Inferred genealogies' tab provides a simple tool to post-process offspring genealogies inferred using NMπ (see 5.1 for details about this output file). In order to post-process the output file containing inferred genealogies, first the appropriate data input file must be selected in the tab 'Data'. Then, the file containing genealogies can be selected. As a result, wNMpi creates a table which can be copy-

pasted (as a tab-delimited text) into another file or software (e.g. Excel). Contents of the table are described in Table 2.

Table 2. Columns appearing in the table created after selecting output file containing genealogies (in order of appearance).

Name	Description
Prog	Identity number of progeny
Mom	Identity number of assigned mother (-1 denotes unsampled individual)
Dist2Prog	Distance between assigned mother and offspring
Femal	Femaleness of assigned mother
Phen_n	Value of the n-th phenotypic character for assigned mother
Dad	Identity number of assigned father (-1 denotes unsampled individual)
Dist2Mom	Distance between assigned father and assigned mother
Dist2Prog	Distance between assigned father and offspring
Femal	Femaleness of the assigned father
Phen_n	Value of the n-th phenotypic character for assigned father
Prob	Posterior probability of genealogy
Mism	Number of genotypic mismatches between parents and progeny
Ambiguous	(Optional) True/False value showing whether genealogy is ambiguous or not in respect to assignment of sex functions

Various filtering options are provided. The most important filter, and primarily applied to genealogies, is the 'minimum threshold probability for genealogy'. Any value larger than zero results in filtering out genealogies having the maximum posterior probabilities (`Pr1`; see 5.1) lower than the threshold value. If any of the following is checked: 'Filter out seed immigrants'/'Filter out pollen immigrants'/'Filter out selfed progeny', then all genealogies representing a given class are filtered out.

Although sex-specific assignments are generally possible (and likely a result) under the neighborhood model, some genealogies may still be uncertain (ambiguous) in respect to sex function of assigned parental individuals. The filter 'Include ambiguous assignments' allows identifying such ambiguous assignments. Formally, a genealogy is called ambiguous if two the most likely genealogies of the offspring consist of the same individuals but their sex functions are reversed. If 'Include ambiguous assignments' is checked then such genealogies are also included (and specified by a value of true vs. false in the additional column 'Ambiguous'). For ambiguous genealogies, the probability is given as the sum of `Pr1` and `Pr2` (see 5.1).

6.5 'Results summary' tab

This section provides functionality for the Maximum Likelihood estimation only. Results of the Bayesian estimation under the Hierarchical Neighborhood Model can be read directly from the output file.

Using NMπ, one can estimate parameters under many alternative models. For example, one can be interested in how differently shaped dispersal kernels influence immigration rates and selection gradients. If the results of multiple analyses are saved to differently named output files (see section 5.1 and 6.1), the 'Results summary' tab can be used to extract the likelihood value as well as parameter estimates from the output file(s) generated by NMπ. Using this option, one can easily inspect multiple output files in order to identify changes in the likelihood value (or AIC) and parameter estimates under

different models. Different parameter types can be filtered out using check boxes. The content of the resulting table can be copy-pasted using the context (right-click) menu.

Note that multiple files can be selected in the open file dialog box using a combination of [Shift] or [CTRL] key and the right click.

6.6 'Preferences' tab

'Preferences' tab is dedicated to various settings, including the number of threads (cores) to use or the path for NMpi.exe file.

7 Estimation strategies

The sections 7.1-7.3 explain how to get reasonable estimates if the Maximum Likelihood estimation (the standard Neighborhood model) is chosen. Information related to the Bayesian estimation (the hierarchical Neighborhood model) is given in the section 7.4.

7.1 How to get estimates of parameters under MLE (standard model)

Because the neighborhood model is a complex probability model, simultaneous estimation of parameters is generally not trivial. Usually, it is rather unlikely to get estimates of all parameters of interest after a single run. This is because the estimation is based on the numerical algorithm (Newton-Raphson), which is very efficient in terms of convergence speed but also highly sensitive to starting values. If analysis is initialized with parameter values which are far from the maximum arguments (true values), the algorithm can behave wildly and often can fail to find the maximum (to reach convergence). Since version 1.2, NMπ offers the additional initial searching for optimal parameter values. To enable this function, one needs to set `max_piter` to a non-zero integer (in the Windows interface it is set to 50 by default). However, this function may fail from time to time so there might be a need for alternative approach.

One remedy for this is to estimate parameters using a step-wise approach. My own experience shows that it is highly advisable to first get the estimates of seed (for dispersed data only) and pollen immigration and mistyping error rates, because these parameters are critical for the behavior of the estimation algorithm. To do this, for the first run, immigration and error rates must be set to non-zero values as well as they need to be set estimable. Usually, estimates are very close to the final estimates, when all significant factors are included into the model. It is worth mentioning that the model with immigration parameters and genotyping error rates only represents quasi-null mating model, because mating within a study population is assumed to occur at random. It can be used as a reference model for studying the effects of putative determinants of mating process, including non-random dispersal and phenotypic effects.

Further strategy relies on available data. Again, my own experience allows me to suggest that performing an analysis similar to a regression analysis, namely a soft-max regression analysis, is often very efficient. If completed, analysis can be continued with some modifications of dispersal kernel function.

The soft-max regression is a special case of multinomial logistic regression, where individual probabilities are linked with presumed factors, i.e. characteristics measured for each individual. In fact, the original neighborhood model can be treated as the soft-max regression for individual male

reproductive success rates (or proportion of sired seeds among the total seed sample) (Adams and Birkes 1991). Individual maternity and paternity probabilities are linked with explanatory variables, such as distance between mates or adult and offspring, adult size etc. It is worth mentioning that the soft-max regression uses an exponential function for all explanatory variables. In the case of non-distance variables, this means that explanatory variables are assumed to have a multiplicative effect on reproductive success (Smouse et al. 1999). When a distance is used as explanatory variable, the soft-max regression implies that dispersal follows an exponential dispersal kernel (Burczyk and Koralewski 2005). Finally, when a direction is used as explanatory variable, the soft-max regression implies that distribution of direction follows the von Misses distribution.

In order to perform the soft-max regression, dispersal kernels must be set to 'Exponential-power' and the shape parameters must be fixed at 1 (treat as not estimable). Distance parameter must be used as a reciprocal of mean dispersal distance (see 4.1) and set to possibly small value (0.01 can work in many examples). If additional phenotypic characters are included, each must be set estimable and initialized at a possibly low value (0.05 works in many cases, given that phenotypic values are given as standardized measurements[1]).

7.2 Comparisons between models

The neighborhood model can contain different combinations of parameters. Sometimes it is worthwhile to compare alternative models more formally. The simplest way is to perform the likelihood ratio test, which is based on the maximum values of the likelihood function for competing models, i.e.

$$LR_{ij} = \frac{L_i}{L_j},$$

where L_i and L_j is the maximum likelihood for the i-th (null) and j-th (alternative) model, respectively. According to the Wilks' theorem, $-2 \times \ln LR_{ij}$ has the asymptotic Chi-Square distribution with degrees of freedom (df) equal to the difference in dimensionality (a number of estimable parameters) between the i-th and j-th model. In practice, log-likelihoods are used so that $-2 \times \ln LR_{ij} = -2(\ln L_i - \ln L_j)$. Thus, the difference in log-likelihoods between competing models of at least 1.92 when df=1 can be interpreted that the difference between models is statistically significant.

Important note:
Models for different data set **cannot** be compared with the likelihood ratio test. Also, if models for the same data differ in the value of `neighborhood` the likelihood ratio test **is not valid** in general, because `neighborhood` defines number of individuals included into analysis (data underlying likelihood values are not the same). The same restrictions apply to Akaike Information Criterion.

[1] Standardized measurement for the i-th individual and the j-th trait is computed as $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$, where x_{ij} is the measurement in original units, e.g. a diameter of the i-th adult in cm, $\bar{x}_j$ is the sample average of the j-th trait and s_j is the sample standard deviation.

Another way to compare models is to use the Akaike Information Criterion (AIC), which penalizes for dimensionality and allows to avoid over-parameterized models. For the i-th model with k estimable parameters AIC is computed as

$$AIC_i = -2 \times \ln L + 2 \times k.$$

Among M competing models, the model with the lowest AIC (AIC_{min}) performs best. Based on AIC it is also possible to estimate model weights for a series of M competing models. The weight of the i-th model is computed as

$$w_i = \frac{\exp\left(-\frac{AIC_i - AIC_{min}}{2}\right)}{\sum_j^M \exp\left(-\frac{AIC_j - AIC_{min}}{2}\right)}$$

Weights sum up to 1 and can be used to estimate multi-model means of parameters. The advantage of multi-model inference about parameter values is that model uncertainty is taken into account. The mean across M models is computed as a weighted average, i.e.

$$\bar{\theta} = \sum_i^M \theta_i w_i.$$

Note, however, that standard error of $\bar{\theta}$ must take into account both the variances within models and between models. Burnham and Anderson (2002) showed that the variance of $\bar{\theta}$ can be computed as

$$V_{\bar{\theta}} = \left[\sum_i^M w_i \sqrt{V_{\theta_i} + (\theta_i - \bar{\theta})^2}\right]^2,$$

so that the variance of a multi-model mean of a given parameter $\bar{\theta}$ can be computed substituting square of the estimate of standard error given in the output file under the i-th model (see 5.1) for V_{θ_i}. Then, square root of $V_{\bar{\theta}}$ is the unconditional[2] standard error of $\bar{\theta}$.

7.3 Dry parentage assignment

A very interesting option is a possibility to perform 'dry' parentage assignment, i.e. to infer genealogies without any estimation of the neighbourhood parameters. The procedure is based on the same reasoning as in the full estimation procedure, but all the parameters are set constant (`incl` = 0) while starting the analysis. In this case NMπ will not estimate parameters, obviously. However, it will perform parentage assignment and create a '.par' output file, as normally. This approach is useful if one is interested in fast screening of data for robust genealogies (e.g. with probabilities > 0.9). For example, if seeds were collected beneath selected individuals, setting `ds` to an average distance of collected seeds from their putative mothers (with `dist_typ`=0) allows for quite robust reconstruction of seed families. However, it is recommended to set some portion of selfing (if biologically reasonable) and seed/pollen immigration as well as non-zero error rates (1-2% may be usually okay).

[2] Unconditional, because such estimate of standard error does not depend on any particular model

7.4 Estimation strategies if the Bayesian approach is chosen (the hierarchical model)

Important note:
Generally, if nmpi2.exe is run on a personal computer (laptop or desktop PC with a several-cores processor), it is not advisable to run the Bayesian approach with genotyping errors set as estimable parameters (i.e. `incl_mtyp = 1`, see the section 4.2). Under such settings, the analysis will be discouragingly slow for most real-world data (see Chybicki 2023 for some discussion on this topic). If one has no access to a server-type computer (with 20+ cores), they can still run the Bayesian analysis that properly accounts for genotyping errors. For this purpose, however, error rates need to be treated as known constants (i.e. `incl_mtyp = 0`). Error rates can be guessed (rather a bad idea!) or determined based on re-genotyping (if possible). My personal favorite approach is to estimate error rates using nmpi2.exe with the ML analysis. For this purpose, I would recommend rather a simple neighborhood model, where the following parameters are included: `ms`, `s`, `mp`, `ds`, `dp`, and a series of `mtyp`, all being estimable. The estimated values of `mtyp` (saved in the output file 'errors.out') can then be used to feed the Bayesian analysis.

The Bayesian approach implemented in the NMπ software is generally time-consuming, even if a multi-core CPU architecture is used. Therefore, the estimation procedure was designed to run in a fully automatic mode. However, since the method uses partly the Metropolis-Hastings algorithm, there might be potential issues with convergence of the simulated Markov Chain. Even if the number of MCMC iterations is large (100,000 is recommended), convergence problems may still arise as a result of bad mixing (too high or too low) across the iterations due to bad proposal distributions. Proposal distributions are automatically pilot-tuned during the initial run (20,000 iterations) and usually they are adjusted correctly. However, the implemented procedure may fail occasionally. One way to guarantee good mixing property is to check acceptance rates of parameters across MCMC. They are shown at the very end of the result file (*.hnm). If the final acceptance rates are outside the 25-45% interval, the results should be treated with caution.

Another source of uncertainty about estimates is the number of explanatory variables (phenotypic characters) considered in the analysis. The Bayesian method implemented in NMπ attempts to select the most plausible regression model for fecundity using the model selection approach called the Reversible Jump MCMC. The approach aims to generate the posterior distribution of candidate regression models. As a result, the model with the highest probability can be selected as the best explanatory model. However, since the number of candidate regression models increases geometrically with the number of phenotypic characters, the number of MCMC iterations required to explore the distribution of candidate regression models can exceed capacity of computer systems. As a result, neither candidate model may appear to be an obvious candidate for the best model. Under such conditions, the effective roundabout is to perform the analysis in two steps. In the initial step, all phenotypic characters are included in the analysis. Based on posterior inclusion frequencies of phenotypic variables estimated in the initial run, the subset of important explanatory variables can be selected for the final run. A good practice is to nominate only those variables which are characterized

by the posterior inclusion frequency ≥0.5. In the literature such a subset of variables corresponds to the 'median' model. The posterior inclusion frequencies can be found in the result file (*.hnm).

8 How to compile NMπ from the source?

The FORTRAN source code of NMπ is provided together with the other files (nmpi2.f90). The code was written in GNU Fortran and can be successfully compiled under a variety of operating systems (for more information visit the GCC web site) using the GCC compiler (starting from version 4.9). In order to enable parallelized computing the flag '-fopenmp' needs to be added while compiling. Also '-O3' and '-funroll-loops' flags can improve the performance.

9 How to cite NMπ?

The list of publications describing the software or major improvements:

- Chybicki IJ (2018) NMπ - improved re-implementation of NM+, a software for estimating gene dispersal and mating patterns. Molecular Ecology Resources 18:159–168
  (This paper introduces the first version of the software)
- Chybicki IJ, Oleksa A, Dering M (2021) Identification of determinants of pollen donor fecundity using the hierarchical neighborhood model. Molecular Ecology Resources 21: 781-800
  (This paper introduces the hierarchical Neighborhood Model)
- Chybicki (2023) NMπ software update to minimize the risk of false positives among determinants of reproductive success. Molecular Ecology Resources. In press.
  (This paper introduces the second version of the software where the seedling version of the hierarchical neighborhood model is implemented)

10 Final remarks

NMπ is an uncommercial software and may contain bugs, although the author has done his best to check for errors in all routines. Therefore, the user is kindly requested to inform about any erratic behaviour. Although a regular assistance cannot be assured, the user can count on the author's help, if needed. In this case he/she is requested to e-mail at igorchy@ukw.edu.pl. E-mails will be replied as soon as possible.

11 Credits

The software and the manual were improved thanks to suggestions of the following people: Jarek Burczyk, Ela Sandurska, Olivier Hardy. However, errors remain my own.