cembra

The user manual

written by Igor J. Chybicki

Department of Genetics
Kaziemirz Wielki University
Chodkiewicza 30, 85064 Bydgoszcz, Poland

(The last update: 16.12.2014)

Contents

Introduction 1
Preparing data 2
Data import 3
Data export 3
The 'Settings' window 3
Performing the analysis 4
Interpreting the results 5
The difference between coding systems used for nominal variables 6
The over-dispersion model for mating system parameters 7
Software limitations 9
References 9

Introduction

cembra was designed in order to make inferences about the effect of categorical or continuous variables on outcrossing rates. The statistical procedures are based on a logistic regression analysis, where a (normally known) binomial response variable is substituted with an unknown outcrossing rate. The estimation is conducted within a Bayesian framework, using the Markov Chain Monte Carlo approach. The details of the model are described in Chybicki and Dzialuk (2014). In addition, the program offers the routine for estimation of outcrossing and effective selfing rates, with the latter representing the parameter useful for quantification of biparental inbreeding (Chybicki, submitted). This manual attempts to give the information necessary to use the software. However, in case of question, please contact the author at the e-mail: igorchy@ukw.edu.pl.

Preparing data

Important! The software assumes by default that progeny genotypes are fully compatible with maternal genotypes provided in the data file. Eventual incompatibilities will result in the crush as quickly as the program starts to analyze data. In order to relax the compatibility assumption, see 'Software limitation' section.

Generally, data must be provided in a tab- or space-delimited text file. Except for continuous variables, all the values must be integers. The values of continuous variables are assumed to be real numbers, with '.' used as the decimal separator.

In the data file, there are 4 integer numbers in the first row: the number of progeny groups (G), the number of loci (L), the number of nominal predictors (N), the number of continuous predictors (C) of outcrossing or effective selfing rates. In the second and next rows, a list of N + C variable names follow. There must be no empty line(s) between the first (i.e. G L N C) and the second row (i.e. the name of the first variable), so be careful! Then, G blocks follow, each representing a single group of progeny. In the example below, there are 2 progeny groups (families), individuals are genotyped at 3 loci, there are 2 nominal predictors and 1 continuous predictor:

```
2     3     2     1
Nominal_var1
Nominal_var2
Continuous_var1

1     2     40.5
200   200   149   149   342   342
3
198   200   149   149   342   342
200   200   149   149   342   350
200   202   -1    -1    342   348

2     1     53
164   200   147   149   342   376
4
188   200   147   149   374   376
164   164   147   149   342   342
164   200   147   147   342   376
164   200   149   149   342   342
```

Each progeny block starts with the header row, where (N + C) values, for the declared nominal and continuous predictors (in that order!), are given. Then, in the next line, a maternal genotype is given. In the third line, a single integer N is given, equal to the number of progeny in the block. Then, N rows follow, each containing the genotype of a single progeny individual. Missing genotypes (both alleles are assumed to be missing!) are coded with double '-1'.

Note that a single family can be splitted appropriately into several blocks, according to the grouping variable (either nominal or continuous). For example, if one collected seeds from a given plant in two (or more) sampling seasons, then two (or more) progeny blocks would represent different seasons.

However, one must remember that maternal genotype must be provided separately for each block, even if repeated across different blocks.

Note that a scale in which continuous variables are expressed will determine the value of the slope coefficient. If one wants to compare the effect sizes among continuous variables, then the z-standardized values must be provided (e.g. Schielzeth 2010). However, in this case the slopes may be more difficult to interpret in terms of the impact of a change in a variable (in measurement units) on the outcrossing/selfing odds.

Data import

Data formatted in the way described above can be imported using the 'File|Open' command in the main menu. Once data are successfully imported, the basic information is shown in the screen.

If categorical variables are provided, the software performs the chi-squared test for associations between variables (categories). Generally, in a regression analysis, predictors must be uncorrelated. Thus, the results of the test may help in determining which, if any, variables are associated and, in consequence, may be redundant in the analysis. Normally, assuming the maximum likelihood estimation, including redundant variables will result in different results, depending on the order of variables. However, in the case of Bayesian approach used here, both (or more) redundant variables may be not significant, even if each single variable is significantly associated with outcrossing rates. However, the last property depends strongly on the length of the Markov chain. If too short chains are used, the results may suggest that the slopes for correlated variables behave normally, leading to false significant estimates. Therefore it is recommended to use long chains, following the suggestions below.

Note that the software offers no function to test whether continuous variables reveal associations. In that case, the user must perform the appropriate correlation analysis in advance, using the external software. Note that associations may exist even between nominal and continuous variables. In this case, there is no proper correlation measure to be used. However, ANOVA can be helpful in this case, given that continuous variables are normally distributed.

Data export

The imported data can be exported to different file formats, incl. MLTR (Ritland 2002), MSF (Chybicki 2013), POLDISP (Robledo-Arnuncio et al. 2007) and SpaGeDi (Hardy and Vekemans 2002).

[to complete]

The 'Settings' window

The 'Analysis|Settings' menu command allows to set-up the analysis. The user can choose between single- and multi-locus outcrossing model. Once multi-locus model is chosen, the user can choose to include biparental inbreeding (effective selfing) or not. Also, the user can choose between the regression-based and the over-dispersion model for outcrossing and/or effective selfing rates. The latter is described in a separate section below. Furthermore, the length of MCMC run ('number of samples', NS), the length of burn-in ('Number of disregarded samples', NB) and thinning can be set

(NT). Those values should be chosen carefully, because they determine the quality of the estimates. As generally known, NS should be large enough in order to satisfy convergence of the Markov chain. The default value of 100 000 should be good enough for the rough estimation of the parameters. However, 1000 000 samples should be used in most cases, if one wants to provide highly repeatable estimates. The value of NB does not need to be very high, because the Markov chain resulting from the algorithm tends to mix quite quickly. As a rule of thumb I would suggest to set NB to be no more than 10% of NS, which is likely much more than required in most cases. Recently, thinning was shown to be unnecessary in most cases (Link and Eaton 2010). However, thinning is still helpful in managing very long chains, especially in the context of the required storage capacity. In order to keep a reasonable size of the output file, I suggest to set NT to a value to get no more than 20 000 stored samples, i.e. NS/NT < 20 000. On the other hand, NS/NT should be also not less than 1000; otherwise the sample from the posterior distribution may be too small to compute precisely the quantiles.

Note that the software does not check if NS > NB or NS > NT. Please assure that those conditions are met!

If the regression-based model is chosen, further settings are available, including the coding system (for nominal variables), the reference levels (for nominal variables) and the variance of the prior distribution for scale coefficients (10 is a default value). Note that coding system changes the interpretation of slope coefficients, having no effect on the overall fit and model quality. The difference between coding systems is explained in more detail in a separate section below.

The variance of the prior distribution for slope coefficients is used to determine the shape of the prior distribution. The default value should be okay for most cases. However, the user may want to experiment a bit (probably by increasing the value), in order to determine if this value changes the results. Generally, the larger variance the less informative prior.

Using ‘Include/exclude effects’ list, the user can choose which, if any, factors are included in the model. If no variable is selected, the resulting model will contain the constant term only (the null model).

If the over-dispersion model is chosen, the user can choose between ‘the beta-binomial model for overdispersion’ or the null model.

To confirm all the options, one clicks the ‘OK’ button. Note that the settings cannot be changed during the analysis.

Performing the analysis

To run the analysis, the ‘Analysis|Run’ command is used. Usually, a single run takes from several minutes to several hours, depending on data and the length of the Markov chain. During the analysis, the software shows the approximate time left to the end. The analysis can be cancelled using the ‘Analysis|Cancel’ command. However, the results will not be shown. During the analysis, the samples (taken from the posterior distribution) are written automatically to the output file ‘mcmc.out’ created in the same directory as the input file. Also, at the end of the run, the second output file is created, ‘ind_t.txt’, containing the individual outcrossing rates. These values are in fact the individual

posterior probabilities (i.e. estimated separately for every progeny individual), that a given progeny was a result of outcrossing.

Note that both 'mcmc.out. and 'ind_t.txt' files are over-written automatically (without prompting!) during every new run. Therefore, if one wants to keep those files, I suggest to re-name them or move them into another directory.

Note that the output files are tab-delimited (with '.' as the decimal separator) and can be easily copy/pasted into a spreadsheet software (e.g. Excel) or imported into any statistical package (e.g. R).

Immediately after the end of run, the results are shown in the screen. See the next section for the interpretation of the results.

Interpreting the results

Once the analysis is completed, the results are shown in the screen in a form of tab-delimited text (copy/paste into a spreadsheet may be helpful in order to interpret the results). First, the information is shown whether multi- or single-locus outcrossing model was used. Then, if the regression model is used, detailed information is given, which variables were included (1) or excluded (0) and which levels (nominal variables only) were used as a reference in the analysis. Also, the coding system is specified. Then, the information about the log-likelihood of the model for posterior averages, the average and variance of the log-likelihood across MCMC is given. It is followed by DIC-related indices: DBar – the average deviance, DHat – the deviance for the posterior averages, pD – the effective number of parameters and DIC – the deviance information criterion. Finally, the information about the value of the variance of prior distribution for slope coefficients is shown. The interpretation of DIC is briefly discussed in a separate paragraph below.

The final estimates of the parameters are shown in a separate table entitled 'Summary of posterior marginal distributions'. Here, posterior mode, mean and some quantiles are given, including 2.5%, 5%, 25%, 50%, 75%, 95% and 97.5% quantile of the posterior distribution. Note that 50% quantile corresponds to median. Also, 2.5% and 97.5% quantiles can be used as limits of 95% (equal tails) credible interval. Also, the limits of 95% highest posterior density interval are shown, i.e. HPDl(95%) and HPDh(95%). If the regression model is used, alpha (constant term), beta (slopes for nominal variables) and gamma (slopes for continuous variables) parameters of the regression function are shown together with the estimated individual outcrossing rates for family groups. If the over-dispersion model is used, the table shows the estimates for the hyper-parameters of the beta-binomial model (mt, yt, me and my) together with individual outcrossing rates for family groups. The hyper-parameter mt and yt (me and ye) is the mean and dispersion of the beta prior. The former may be referred to as the population mean outcrossing rate (me is the population mean effective selfing rate). The latter is used to measure over-dispersion. Details are described in the section below.

Finally, the table with the estimated pollen allele frequencies is shown. However, although these estimates are fully Bayesian, posterior averages are only extracted.

Using DIC

DBar is a measure of model fit (or unfit). For any two model with the same (effective) number of parameters, one with the smaller DBar fits better to data. Generally, the more parameters in the

model, the better fit. So, in order to account for over-parameterization, one needs to penalize for the number of parameters. This is done using DIC, which is computed as DIC = DBar + pD (model fit penalized by the number of parameters). Here the analogy to the Akaike Information Criterion (AIC) is clearly seen. The effective number of parameters, pD, can be interpreted as a number of parameters in the model unconstrained by the prior distribution.

Because DIC is a large sample Bayesian analog of AIC, in order to compare any two model one may follow the rules for AIC. After Burnham and Anderson (2002), if the DIC difference between the i-th model and the best model (with the lowest DIC) is between 0-2, the i-th model still has a substantial support, between 4-7 considerably less support and >10 essentially no support. When two competing models have a substantial support, according to maximum parsimony approach, I would prefer that with the lower pD.

Note that, although potentially informative about model weights, DIC was not proved as a proper means for Bayesian model averaging. Also, in order to assure stable DIC, the Markov Chain needs to converge. In most cases, the default number of samples (100 000) is not enough. See the 'The 'Settings' window' section above.

The difference between coding systems used for nominal variables

To perform a regression analysis for nominal predictors, the software allows to choose between two coding systems: *Deviation from means* and *Reference cell*. First let us clarify terminology used in this section. A nominal predictor is assumed to have K levels. For example, seeds may be grouped according to population size, when 3 types are distinguished (small, moderate, large). In this case, a population size represents a single nominal predictor with 3 levels. Generally, i.e. regardless the coding system, K - 1 code (design) variables are generated for K levels. In consequence, K - 1 unredundant slope coefficients are required to map the association between a nominal variable and a response variable. Thus, if more nominal explanatory variables are provided, for every predictor (K – 1) slope coefficients are required.

Specification of the code variables for a nominal predictor and the corresponding regression function terms.

Coding system	Level	Code variables		Regression function term
		x_1	x_2	
Deviation from means				
	1	1	0	$\beta_1 x_1 + \beta_2 x_2 = \beta_1$
	2	0	1	$\beta_1 x_1 + \beta_2 x_2 = \beta_2$
	3	-1	-1	$\beta_1 x_1 + \beta_2 x_2 = -(\beta_1 + \beta_2)$
Reference cel (here level 3)				
	1	1	0	$\beta_1 x_1 + \beta_2 x_2 = \beta_1$
	2	0	1	$\beta_1 x_1 + \beta_2 x_2 = \beta_2$
	3	0	0	$\beta_1 x_1 + \beta_2 x_2 = 0$

Two coding systems are shown in the table above. One may see that using *Deviation from means* for a variable with K levels results in K regression terms, of which K – 1 are actually estimable slope coefficients, while the remaining one is always a negative sum of these K – 1 coefficients. However, it does not matter, which level is chosen to be 'reference' in this case. In the example shown in the table, the values will remain unchanged if, instead of the third level, the first one is chosen.

In the case of *Reference cell* coding (or *Dummy coding*), one needs to choose the reference level (level 3 in the table), to which all the other levels will be compared. Sometimes, especially in the case of K = 2, it may be natural to think of one level as to be the 'reference'. For example, if one wants to test for the effect of experimental stimulation of flowering on outcrossing, two groups of progeny can be distinguished (stimulated vs. control). In this case the control group may be a natural reference level. However, such a natural classification is not possible sometimes. Then, *Deviation from means* may be preferred.

Basically, the difference between the two coding systems is in the interpretation of resulting slope coefficients. In the case of *Deviation from means*, the slopes represent the difference in outcrossing/selfing odds between a particular level and a geometric mean of odds for a nominal variable (i.e. across all the levels). In this case, the value of zero is meaningful in respect to the overall insignificance of a particular level. In the case of *Reference cell*, the slopes represent the odds ratio for a given level to the reference level. In other words, the slopes inform how the outcrossing/selfing odds change in a given level as compared with the reference level. In this case, the value of zero informs that there is no difference between a given level and the reference level.

Note that, unlike *Deviation from means*, in the case of *Reference cell* coding, changing a reference level leads generally to different estimates (and interpretation) of the slope coefficients. Nonetheless, the overall model fit remains unchanged.

The over-dispersion model for mating system parameters

The software can be also used to verified whether the outcrossing and effective selfing rates for individual progeny groups values reveal any extra variability over that expected for a binomial distribution. In other words it allows to verified whether there is any over-dispersion in the individual outcrossing rates, ignoring however the impact of provided variables, if any. The significant deviation from a binomial distribution would imply that there was a significant factor behind the variation in individual outcrossing rates, including such features as the location (population), variable pollen availability or a variable number of lethal alleles.

The approach is generally based on the (mixed mating) probability model (Chybicki, submitted), similar to that used in MLTR (Ritland 2002), in which the probability of the multilocus genotype O_{ij} of the j-th offspring in the i-th maternal family is equal

$$\Pr(O_{ij}) = (1-t_i)\prod_l^L \Pr(O_{ijl}|M_{il}) + t_i \prod_l^L (e_i \Pr(O_{ijl}|M_{il}) + (1-e_i)\Pr(O_{ijl}|M_{il}, \mathbf{P}_l)), \qquad [1]$$

where t_i is the probability that a seed collected from the i-th mother plant was produced through outcrossing, e_i is the probability that a seed collected from the i-th mother was produced through outcross mating between relatives, $\Pr(O_{ijl}|M_{il})$ is the Mendelian probability of the offspring genotype after self-fertilization given the maternal genotype M_{il} at the l-th locus, while

$\Pr(O_{ijl}|M_{il}, \mathbf{P}_l)$ is the Mendelian probability of the offspring genotype after outcrossing given the maternal genotype M_{il} and the background pollen pool $\mathbf{P}_l$ (represented by the array of allele frequencies) at the l-th locus. The likelihood function of the model is

$L(\mathbf{O}; \mathbf{t}, \mathbf{P}) = \prod_i \prod_j \Pr(O_{ij})$. [2]

In order to estimate over-dispersion, t_i and e_i are assumed to follow a beta distribution:

$f(t; \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$, [3]

where α and β are the parameters such that the expected value equals $\mu = \frac{\alpha}{\alpha+\beta}$, while the variance equals $\sigma^2 = \mu(1-\mu)\frac{1}{1+\alpha+\beta}$, while θ refers to the parameter of the model (either t or e). The far right-hand side term $\frac{1}{1+\alpha+\beta}$, sometimes called the dispersion parameter γ, is relevant to study over-dispersion. Therefore, the model uses a Beta distribution expressed in the equivalent form with μ and γ parameters, using the following substitutions $\alpha = \frac{\mu(1-\gamma)}{\gamma}$ and $\beta = \frac{(1-\mu)(1-\gamma)}{\gamma}$. The parameters μ and γ take values within (0,1). In the case of γ, the asymptotic 0 means that there is no over-dispersion (e.g. the number of outcrossed progeny within families follows the binomial distribution with the same parameter), while asymptotic 1 means that there is the extreme over-dispersion (e.g. the number of outcrossed progeny within each family follows the binomial distribution with the unique parameter). Both parameters are set to be estimable along with the parameters of the likelihood function [2].

The parameters are estimated using the Gibbs sampler (a class of Markov Chain Monte Carlo or MCMC algorithm). A uniform Dirichlet distribution is taken as a prior for $\mathbf{P}_l$ (i.e. the vector of allele frequencies at the l-th locus), while for outcrossing and effective selfing rates the beta distribution is taken with hyper-parameters μ and γ. In the case of μ and γ, the uniform and the (improper) distribution proportional to 1 / γ are used, respectively. The latter is chosen in order to express the conservative null hypothesis, that there is no over-dispersion in individual outcrossing rates (i.e. γ is close to 0). Although there is a boundary issue (i.e. $\gamma \in (0,1)$), which precludes performing a formal test whether γ = 0, given the conservative prior taken for γ, the Bayesian confidence interval can inform about significant over-dispersion. As a cross-validation for this procedure, one may performed the Bayesian model comparison to verify if the hierarchical ("over-dispersion") model better fits to data than the null model. For this purpose, an additional analysis based on the null model must be performed. In this case, no extra variation in outcrossing rates is assumed (i.e. γ = 0). Conceptually, the estimation algorithm is the same, except for the parameterization of outcrossing and effective selfing rates. In this case a series of t_i and/or e_i is replaced with a single (mean) rate, for which a uniform beta distribution is taken as a prior.

Using DIC, one may compare competing mating models, with differences defined in terms of effective selfing and outcrossing rates. For example, in order to see whether effective selfing (biparental inbreeding) contributes to the observed mating pattern, one may compare models with and without effective selfing. Note that single-locus outcrossing model is an equivalent of the multi-locus mating model, in which there is no self-fertilization (t = 1). Hence, the single-locus outcrossing mode can be chosen to see whether any self-fertilization contributes to the observed mating pattern.

Software limitations

There is no limit *a priori* for a number of families, individuals, loci etc. Also, there is no limit for a number of explanatory variables (predictors). However, having a very large sample one may experience increased analysis time.

By default, there is no possibility for any genotypic incompatibility between progeny and a mother plant. However, if one can assume that scoring problems resemble K-Alleles Model (i.e. due to mutation or scoring mistake, an allele can be substituted with any other allele, drawn from K alleles in total at a given locus), then the **average-over-loci** rate of genotyping error (epsilon, taking values between 0 and 1) can be set using 'Analysis|Set epsilon (…)' command in the main menu. Note that the current setting is shown in brackets. If epsilon is set to nonzero value, incompatibility between progeny and mother plants is no longer an issue. Although the implemented procedure works fine in most cases, it is strongly recommended to use reasonable epsilon value, e.g. estimated experimentally through re-genotyping.

Also, despite that incomplete maternal genotypes are allowed (i.e. with missing single-locus genotypes), the gaps are not reconstructed from progeny. Thus, generally it is not recommended to put many gaps in maternal genotypes, because data use may be suboptimal. If one has problems with providing maternal genotypes, it is recommended to analyze data with MLTR (Ritland 2002) or MSF (the other software developed by the author, which accounts simultaneously for typing errors and the presence of unrelated individuals in maternal families) in advance, in order to estimate maternal genotypes. Another limitation is that there is no option to choose different coding systems independently for each nominal variable. Also, there is no possibility for missing values in the case of predictor variables. However, in that case the user can substitute missing data with appropriate mean values. Some of these limitations will be relaxed in future versions.

References

Burnham KP, Anderson DR (2002). Model selection and multimodel inference: a practical information-theoretic approach. 2nd Edition. Springer-Verlag, 488 pp.

Chybicki IJ. Quantifying biparental inbreeding based on progeny arrays: old methods vs. new concepts. *Submitted.*

Chybicki IJ (2013) Note on the applicability of the F-model in analysis of pollen pool heterogeneity. J Hered 104:578-585.

Chybicki IJ, Burczyk J (2013) Seeing the forest through the trees: comprehensive inference on individual mating patterns in the mixed stand of *Quercus robur* and *Q. petraea*. Ann Bot 112:561-574.

Chybicki IJ, Dzialuk A (2014). Bayesian approach reveals confounding effects of population size and seasonality on outcrossing rates in a fragmented subalpine conifer. Tree Genet Genomes 10:1723-1737.

Hardy O, Vekemans X (2002) SPAGeDi: a versatile computer program to analyse spatial genetic structure at the individual or population levels. Mol Ecol Notes 2:618–620.

Link WA, Eaton MJ (2010) On thinning of chains in MCMC. Methods Ecol Evol 3:112-115.

Ritland K (2002) Extensions of models for the estimation of mating systems using n independent loci. Heredity 88:221-228.

Robledo-Arnuncio JJ, Austerlitz F, Smouse PE (2007) POLDISP: A software package for indirect estimation of contemporary pollen dispersal. Mol Ecol Notes 7:763-766.

Schielzeth H (2010) Simple means to improve the interpretability of regression coefficients. Methods Ecol Evol 1:103–111.